Cayley Differential Unitary Space-Time Codes

BABAK HASSIBI* BERTRAND M. HOCHWALD**

*Department of Electrical Engineering
California Institute of Technology, Pasadena, CA 91125
e-mail: hassibi@systems.caltech.edu ‡

**Bell Laboratories, Lucent Technologies
600 Mountain Avenue, Murray Hill, NJ 07974
e-mail: hochwald@lucent.com

August 31, 2001

Abstract

One method for communicating with multiple antennas is to encode the transmitted data differentially using unitary matrices at the transmitter, and to decode differentially without knowing the channel coefficients at the receiver. Since channel knowledge is not required at the receiver, differential schemes are ideal for use on wireless links where channel tracking is undesirable or infeasible, either because of rapid changes in the channel characteristics or because of limited system resources. Although this basic principle is well understood, it is not known how to generate good-performing constellations of unitary matrices, for any number of transmit and receive antennas and for any rate. This is especially true at high rates where the constellations must be rapidly encoded and decoded.

We propose a class of *Cayley codes* that works with any number of antennas, and has efficient encoding and decoding at any rate. The codes are named for their use of the Cayley transform, which maps the highly nonlinear Stiefel manifold of unitary matrices to the linear space of skew-Hermitian matrices. This transformation leads to a simple linear constellation structure in the Cayley transform domain and to an information-theoretic design criterion based on emulating a Cauchy random matrix. Moreover, the resulting Cayley codes allow polynomial-time near-maximum-likelihood decoding based on either successive nulling/cancelling or sphere decoding. Simulations show that the Cayley codes allow efficient and effective high-rate data transmission in multi-antenna communication systems without knowing the channel.

Index Terms—Wireless communications, BLAST, differential modulation, unitary space-time codes, transmit diversity, receive diversity, fading channels, Cayley transforms, Cauchy random matrices

‡Prof. Hassibi was with Bell Laboratories, Lucent Technologies, Murray Hill, NJ.

1 Introduction and Model

Although reliable mobile wireless transmission of video, data, and speech at high rates to many users will be an important part of future telecommunications systems, there is considerable uncertainty as to what technologies will achieve this goal. One way to get high rates on a scattering-rich wireless channel is to use multiple transmit and/or receive antennas [1, 2]. Many of the practical schemes that achieve these high rates, such as BLAST (Bell Labs Layered Space-Time) [1], require the propagation environment or channel to be known to the receiver. A variety of design techniques for space-time transmission schemes when the receiver knows the channel have been developed (see, e.g., [3], [4], [5], [6] and the references therein).

In practice, knowledge of the channel is often obtained via training: known signals are periodically transmitted for the receiver to learn the channel, and the channel parameters are tracked (using decision-feedback or automatic-gain-control (AGC)) in between the transmission of the training signals. However, it is not always feasible or advantageous to use training-based schemes, especially when many antennas are used or either end of the link is moving so fast that the channel is changing very rapidly. As the number of transmit antennas grows, the training interval for learning the channel must grow proportionately [7], [8], and the number of pilot signals used to track the channel must also grow. Given a restriction on total pilot or training power, we must allocate less power per antenna with every added antenna. Moreover, schemes such as decision-feedback and AGC can become increasingly complex and prone to error when the number of transmit/receive antennas is large since there are many more channel parameters to adjust. Finally, instability in local oscillators and phase-lock devices and inaccurate knowledge of Doppler shifts, which may be different for each antenna, may also limit channel tracking ability at the receiver.

Hence, there is much interest in space-time transmission schemes that do not require either the transmitter or receiver to know the channel. Some information-theoretic calculations with a channel that changes in a block-fading manner appear in [9], [10], [11], and [12] that suggest that high capacities with multiple antennas are achievable with no channel information if the channel does not change too rapidly. How rapidly the channel may change is not completely clear. For the purposes of this paper it suffices to assume that the channel has a coherence interval (defined to be the number of samples at the sampling rate during which the channel is approximately constant) that is at least twice the number of transmit antennas.

Coding and design criteria for the unknown multi-antenna channel were originally developed in [10], and many design techniques have since been developed that offer reasonable data rates [13], [14], [15], [16], [17]. The technique of [13], while useful for arbitrary numbers of transmit and receive antennas, suffers from complexity difficulties as the number of transmit antennas or data rate grows.

A standard method used to combat fading in single-antenna wireless channels is differential phase-shift keying (DPSK) [18]. In DPSK, the transmitted signals are unit-modulus (typically chosen from a m-PSK constellation), information is encoded differentially on the phase of the transmitted signal, and as long as the phase of the channel remains approximately constant over two consecutive channel uses, the receiver can decode the data without having to know the channel coefficient. Differential techniques for multi-antenna communications have been proposed in [14], [15] and [16], where, as long as the channel is approximately constant in consecutive uses, the receiver can decode the data without having to know the channel. The general differential techniques proposed in [14] and [15] are shown to have good performance when the constellation of matrices used for transmission forms a group under matrix multiplication [19], which also leads to simple decoding rules [20]. However, the number of groups available is rather limited, and the groups do not lend themselves to very high rates (such as tens of bits/sec/Hz) with many antennas. The technique of [16] is based on orthogonal designs, and therefore has simple encoding/decoding and works well when there are two transmit and one receive antenna, but suffers otherwise from performance penalties [6] at very high rates.

We seek a signaling scheme that fits within the framework of [14] but can handle any combination of transmit and receive antennas and any rate. The general design problem for differential transmission, for rate R (in bits/channel use) with M transmit antennas, is to find a constellation of unitary matrices $\mathcal{V} = \{V_0, \ldots, V_{L-1}\}$, with $L = 2^{RM}$, such that $|\det(V_\ell - V_{\ell'})|$ is as large as possible for all $\ell' \neq \ell$. In its full generality, this is an intractable problem since the objective criterion and search spaces are both highly non-convex and the size of the problem is exponentially large in the rate and number of antennas. In [14] and [15] it is shown that there are various simplifications and practical advantages if the set $\mathcal{V}$ forms a group: 1) matrices never have to be explicitly multiplied before transmission; 2) the transmitted matrix is always a member of the constellation. Groups that satisfy the design criterion, i.e., that have nonzero $|\det(V_\ell - V_{\ell'})|$ for all $\ell' \neq \ell$, are referred to as fixed-point-free (fpf) groups.

In [19] all finite fpf groups are classified (see also [21] for L an integer power of two) and many of these are shown to have excellent performance. Nevertheless, the number of finite fpf groups is limited, and good performance is hard to achieve for very high rates and for large numbers of transmit antennas. The infinite fpf Lie groups are classified in [22], where it is shown that there are only two possibilities: $U(1)$ and $SU(2)$—the unit-modulus scalars of single-antenna differential modulation, and the two transmit-antenna orthogonal designs of Alamouti [4]. Therefore, [19] also considers the design of matrices for differential transmission that *do not* form a group; however, the nongroup techniques in [19] do not necessarily lend themselves to simple decoding, and constellation design at very high rates is difficult.

The two mentioned advantages of groups over nongroups are not essential for successful differential transmission and reception. In fact, these advantages are outweighed by our desire for a technique that works for any number of antennas and at any rate; we are therefore forced to consider nongroups. At high rates, where the size of the constellation is very large, the performance of the constellation is determined only in part by the "minimum distance" [14]

$$\zeta = \frac{1}{2} \min_{\ell \neq \ell'} |\det(V_\ell - V_{\ell'})|^{\frac{1}{M}}. \tag{1}$$

Perhaps more important is the general statistical structure of the constellation. At high rates, the structure should statistically emulate the capacity-achieving input distribution.

Part of the difficulty of designing large constellations of unitary matrices is the lack of simple parameterizations of these matrices. To keep the transmitter and receiver complexity low in multiple antenna systems, linear processing is often preferred [23], whereas unitary matrices are often highly nonlinear in their parameters. Part of the success of V-BLAST (vertical-BLAST) for the known channel [24, 25] is its ability to encode and decode rates of tens of bits/sec/Hz by breaking the original data stream into substreams that are transmitted on the individual antennas. The receiver decodes the substreams using a sequence of nulling and cancelling steps. However, the V-BLAST approach does not guarantee unitary matrices and is unsuitable for the differential method.

The Cayley codes we propose also break the data stream into substreams, but instead of transmitting these substreams directly as in V-BLAST, these substreams are used to parameterize the unitary matrices that are transmitted. The codes work with any number of transmit and receive antennas and at any rate. The Cayley codes:

1. Are very simple to encode
2. Can be used for any number of transmit and receive antennas
3. Can be decoded in a variety of ways including simple polynomial-time linear-algebraic techniques such as:
   (a) Successive nulling and cancelling (V-BLAST [24], square-root V-BLAST [26])
   (b) Sphere decoding [27, 28]

4. Are designed with the numbers of both the transmit and receive antennas in mind

5. Satisfy a probabilistic criterion: they maximize an expected distance between matrix pairs

Very brief summary of Cayley Codes

We briefly summarize the general structure of the Cayley codes. To generate a unitary matrix V parameterized by the transmitted data, we break the data stream into Q substreams (we specify Q later) and use these substreams to choose $\alpha_1, \ldots, \alpha_Q$ each from a set $\mathcal{A}$ with r real values (we also have more to say about this set later). We call a rate $R = (Q/M)\log_2 r$ Cayley code one for which V obeys

$$V = (I + iA)^{-1}(I - iA) \tag{2}$$

where

$$A = \sum_{q=1}^{Q} A_q \alpha_q$$

and $A_1, \ldots, A_Q$ are *pre-selected* $M \times M$ complex Hermitian matrices. The matrix V, as given by (2), is referred to as the Cayley transform of iA and, as shown in Section 2, is unitary by construction. The code is completely specified by $A_1, \ldots, A_Q$. Each individual codeword is determined by the scalars $\alpha_1, \ldots, \alpha_Q$.

The performance of a Cayley code depends on the choices of the number of substreams Q, the Hermitian basis matrices $\{A_q\}$, and the set $\mathcal{A}$ from which each α_q is chosen. Roughly speaking, we choose Q so as to maximize the number of independent degrees of freedom observed at the output of the channel. To choose the $\{A_q\}$ we optimize a coding criterion specified in Section 2.5, (equation (37)) that resembles the $|\det(V_\ell - V_{\ell'})|$ criterion given in [14], [15], but is more suitable for the high rates we consider and is amenable to analysis. The optimization is done only once, during code design, and simulations show that it is amenable to gradient-based methods. Finally, for reasons that are specified later, the set $\mathcal{A}$ is chosen as a discrete approximation of a scalar Cauchy random variable.

The Cayley transform (2) is powerful because it generates the unitary matrix V from the Hermitian matrix A, and A is *linear* in the data $\alpha_1, \ldots, \alpha_Q$. In Section 2.4, we show how this leads to simple decoding. Section 3 has several examples of Cayley differential codes and some performance comparisons with existing schemes, and Section 4 concludes the paper. Several mathematical tools and related results used in the paper are developed in the appendices.

We now present a brief summary of the multiple-antenna model and the differential unitary space-time signaling scheme.

1.1 Differential Unitary Space-Time Modulation

In a narrow-band, flat-fading, multi-antenna communication system with M transmit and N receive antennas, the transmitted and received signals are related by

$$x = \sqrt{\rho} s H + v, \tag{3}$$

where $x \in \mathcal{C}^{1\times N}$ denotes the vector of complex received signals during any given channel use, $s \in \mathcal{C}^{1\times M}$ denotes the vector of complex transmitted signals, $H \in \mathcal{C}^{M\times N}$ denotes the channel matrix, and the additive noise $v \in \mathcal{C}^{1\times N}$ is assumed to have independent $\mathcal{CN}(0,1)$ (zero-mean, unit-variance, complex-Gaussian) entries that are temporally white. The channel matrix H is also assumed to have independent $\mathcal{CN}(0,1)$ entries, implying that

$$\mathrm{E}\,\mathrm{tr}\, HH^* = MN.$$

Assuming further that $\mathrm{E}\, ss^* = 1$, and since the random quantities H, s, and v are independent, ρ is the SNR at each receive antenna, independently of M.

The channel is used in blocks of M channel uses. We can then aggregate the transmit row vectors s over these M channel uses into an $M \times M$ matrix S_τ, where $\tau = 0, 1, \ldots$ represents the block channel use. In this setting, the mth column of S_τ denotes what is transmitted on antenna m as a function of time, and the mth row denotes what is transmitted on the M antennas at time m. If we assume that the channel is constant over the M channel uses, the input and output row vectors are related through a common channel so that we may write

$$X_\tau = \sqrt{\rho} S_\tau H + W_\tau, \tag{4}$$

where W_τ and H are $M \times N$ matrices of independent $\mathcal{CN}(0,1)$ random variables and X_τ is the $M \times N$ received complex signal matrix.

In differential unitary space-time modulation [14], [15] the transmitted matrix at block τ satisfies the following so-called fundamental transmission equation

$$S_\tau = V_{z_\tau} S_{\tau-1}, \tag{5}$$

where $z_\tau \in \{0, \ldots, L-1\}$ is the data to be transmitted (We assume $S_0 = I$). Since the channel is used M times, the corresponding transmission rate is $R = (1/M)\log_2 L$. If we further assume that the propagation environment is approximately constant for $2M$ consecutive channel uses, then we may write

$$\begin{aligned} X_\tau = \sqrt{\rho} S_\tau H + W_\tau &= \sqrt{\rho} V_{z_\tau} S_{\tau-1} H + W_\tau \\ &= V_{z_\tau}(X_{\tau-1} - W_{\tau-1}) + W_\tau, \end{aligned}$$

which leads us to the fundamental differential receiver equation

$$X_\tau = V_{z_\tau} X_{\tau-1} + \underbrace{W_\tau - V_{z_\tau} W_{\tau-1}}_{W'_\tau}. \tag{6}$$

Note that the channel matrix H does not appear in the above equation. This implies that, as long as the channel is approximately constant for $2M$ channel uses, differential transmission permits decoding without knowing the fading matrix H.

From (5) it is apparent that the matrices V_ℓ should be unitary, otherwise the product $S_\tau = V_{z_\tau} V_{z_{\tau-1}} \ldots V_{z_1}$ can go to zero, infinity, or both (in different spatial and temporal directions). Moreover, when V_{z_τ} is unitary the additive noise term

$$W_\tau^{'} = W_\tau - V_{z_\tau} W_{\tau-1},$$

is statistically independent of V_{z_τ}. Since the additive noise term W'_τ has independent complex Gaussian entries, the maximum-likelihood (ML) decoder of z_τ is

$$\hat{z}_\tau = \arg \max_{\ell=0,\ldots,L-1} \|X_\tau - V_\ell X_{\tau-1}\|. \tag{7}$$

In [14, 15] it is shown that the pairwise block probability of error (of transmitting V_ℓ and erroneously decoding $V_{\ell'}$) has upper bound

$$P_e \leq \frac{1}{2} \prod_{m=1}^{M} \left[1 + \frac{\rho^2}{4(1+2\rho)} \sigma_m^2(V_\ell - V_{\ell'})\right]^{-N}, \tag{8}$$

where $\sigma_m(\cdot)$ denotes the mth singular value. At high SNR this inequality becomes

$$P_e \lesssim \frac{1}{2}\left(\frac{8}{\rho}\right)^{MN} \cdot \frac{1}{|\det(V_\ell - V_{\ell'})|^{2N}}. \tag{9}$$

Therefore, most design schemes [14, 15, 19, 22] have focused on finding a constellation $\mathcal{V} = \{V_0, \ldots, V_{L-1}\}$ of $L = 2^{MR}$ unitary $M \times M$ matrices that maximizes ζ defined in (1).

In general, the number of unitary $M \times M$ matrices in $\mathcal{V}$ can be quite large. For example, if rate $R = 8$ is desired with $M = 4$ transmit antennas (even larger rates are quite possible as shown later), then the number of matrices is $L = 2^{RM} = 2^{32} \approx 4 \times 10^9$, and the pairwise error between any two signals can be very small. This huge number of signals calls into question the feasibility of computing ζ and lessens its usefulness as a performance criterion. We therefore consider a different, though related, criterion.

The large number of signals also rules out the possibility of decoding via an exhaustive search. For high rates, it is possible to construct a random constellation with some structure [30]. But, again, we have no efficient decoding method.

To design constellations that are huge, effective, and yet still simple, so that they can be decoded in real-time, we briefly examine some parameterizations of unitary matrices and then show how the Cayley transform can be used.

2 Cayley Differential Codes

We first review some properties of the space of all unitary matrices.

2.1 The Stiefel manifold

The space of $M \times M$ complex unitary matrices is referred to as the *Stiefel* manifold. This manifold is highly nonlinear and nonconvex, and can be parameterized by M^2 real free parameters. To see why, note that an arbitrary complex $M \times M$ matrix has $2M^2$ real parameters. Unitarity introduces M^2 real-valued constraints: M constraints to force each column to have unit norm, and $2\frac{M(M-1)}{2} = M^2 - M$ constraints to ensure that the (real and imaginary parts of the) pairwise inner products of any two columns are zero. We now examine some possible parameterizations of the Stiefel manifold.

Parameterization with Givens rotations

A unitary matrix V is often given as the product $V = O_1 D O_2$, where O_1 and O_2 are real orthogonal matrices and D is a diagonal unitary matrix (see, e.g., [31]). A diagonal unitary matrix has diagonal entries with unit modulus: therefore it is described by M real entries, one for the phase of each diagonal entry. An arbitrary real orthogonal matrix, on the other hand, can be expressed as the product of $\frac{M(M-1)}{2}$ Givens (or planar) rotations, one for each of the $\frac{M(M-1)}{2}$ two-dimensional hyperplanes. This implies that we may write

$$V = G_1 G_2 \cdots G_{M(M-1)/2} D G_{M(M-1)/2+1} \cdots G_{M(M-1)},$$

where each G_m is a Givens matrix. Since each Givens rotation is determined by a single real parameter (the angle of rotation), the total number of free variables is M^2, which matches the degrees of freedom in the Stiefel manifold.

It is conceivable that one can use this parameterization to encode data onto the angles of rotation and onto the diagonal phases of D. However, we do not pursue this approach because the parameterization is not one-to-one (one can reorder the Givens rotations, for example), it is highly nonlinear and, most importantly, because we do not know how to decode them in any systematic way.

Parameterization with Householder reflections

A unitary matrix can be written as the product of Householder matrices

$$V = D H_1 H_2 \cdots H_M, \qquad H_m = I - 2\frac{h^{(m)} h^{(m)*}}{\|h^{(m)}\|^2},$$

where D is a diagonal unitary matrix, and each $h^{(m)}$ has the form

$$h^{(m)} = [0, 0, \cdots, 0, 1, h^{(m)}_{m+1}, \cdots, h^{(m)}_{M}]^T.$$

It is not hard to show that this parameterization has M^2 degrees of freedom. However, we also abandon this parameterization since we do not know how to encode or decode the data onto the Householder matrices in an efficient manner.

Parameterization with matrix exponential

The matrix exponential is

$$V = e^{iA},$$

where A is a Hermitian matrix. This method appears propitious because it generates unitary matrices from Hermitian matrices, and it is the matrix generalization of $v = e^{i\theta}$ (used in standard DPSK), where θ is real. The matrix exponential has connections with Lie group theory (if V forms a Lie group, then A forms a real

Lie algebra—see, for example, [32]). An $M \times M$ Hermitian matrix can be parameterized by M^2 free real variables, so the matrix exponential contains the right number of degrees of freedom.

However, the exponential map has the difficulty that it is not one-to-one. This is seen in the scalar case where adding 2π to θ produces the same v. While the scalar difficulty is easily overcome by considering only $a \in [0, 2\pi)$, the equivalent matrix constraint is

$$0 \leq A < 2\pi I,$$

meaning that both A and $2\pi I - A$ are nonnegative definite. Although this constraint is convex, it is nonlinear and we do not know how to sample the space of A's to obtain a constellation of V's. Moreover, unlike the scalar case, the exponential map does not appear to be easily inverted at the receiver when $M > 1$. We therefore do not pursue this approach.

2.2 Parameterization with Cayley transform

The Cayley transform of a complex $M \times M$ matrix Y is defined to be [31, 33]

$$(I_M + Y)^{-1}(I_M - Y), \tag{10}$$

where I_M is the $M \times M$ identity matrix and Y is assumed to have no eigenvalues at -1 so that the inverse exists. (We drop the M subscript on I from now on.) Note that $I - Y$, $I + Y$, $(I - Y)^{-1}$ and $(I + Y)^{-1}$ all commute so there are other equivalent ways to write this transform.

Let A be an $M \times M$ complex Hermitian matrix and consider the Cayley transform of the skew-Hermitian matrix $Y = iA$:

$$V = (I + Y)^{-1}(I - Y) = (I + iA)^{-1}(I - iA). \tag{11}$$

Note that the inverse of $I + iA$ exists because iA has strictly imaginary eigenvalues. The matrix V is unitary because

$$\begin{aligned} VV^* &= (I + iA)^{-1}(I - iA)\left[(I + iA)^{-1}(I - iA)\right]^* \\ &= (I + iA)^{-1}(I - iA)(I + iA)(I - iA)^{-1} \\ &= I, \end{aligned}$$

where we use the fact that A is Hermitian.

Thus, similarly to the matrix exponential, the Cayley transform expresses a unitary matrix as a function of a skew-Hermitian matrix. (Recall that (skew-)Hermitian matrices are described by M^2 real parameters, so that the degrees of freedom match those of the Stiefel manifold.) This parameterization appears promising because it is one-to-one: the Cayley transform can be easily inverted to yield

$$iA = (I + V)^{-1}(I - V), \tag{12}$$

provided that $(I + V)^{-1}$ exists (or, equivalently, V has no eigenvalue at -1). Thus, the Cayley transform and its inverse coincide. The Cayley transform of a unitary matrix (with no eigenvalues at -1) is skew-Hermitian. Indeed, letting $Y = (I + V)^{-1}(I - V)$, with V unitary, we obtain

$$\begin{aligned} Y^* &= (I - V^*)(I + V^*)^{-1} = (V - I)V^*V(V + I)^{-1} \\ &= -(I + V)^{-1}(I - V) = -Y. \end{aligned}$$

We have shown the following result.

Lemma 1 (Cayley Transform and Unitary Matrices). *A matrix with no eigenvalues at -1 is unitary if and only if its Cayley transform is skew-Hermitian.*

Compared with the other parameterizations of unitary matrices, the parameterization with the Cayley transform is not "too nonlinear" (we show why in Section 2.4) and it is one-to-one and easily invertible. The Cayley transform also maps the complicated Stiefel manifold of unitary matrices to the space of skew-Hermitian matrices. Skew-Hermitian matrices are easy to characterize since they form a linear vector space over the reals (the real linear combination of any number of skew-Hermitian matrices is skew-Hermitian). Section 2.4 uses this handy feature for easy encoding and decoding.

2.2.1 Some properties

The Cayley transform (11) is the matrix generalization of the scalar transform

$$v = \frac{1 - ia}{1 + ia}$$

that maps the real line to the unit circle. This map is also called a bilinear map and is often used in complex analysis. The Cayley transform (10) maps matrices with eigenvalues inside the unit circle to matrices with eigenvalues in the right-half-plane. It is therefore often used in systems and control theory to map continuous-time systems to discrete-time systems (since stability is preserved), to map bounded real functions to positive real functions, and contractive systems to passive systems. In the recent references [34, 35], the Cayley transform is used in the numerical solution of differential equations over Lie groups.

The following two results are needed later.

Lemma 2 (Eigenvalues/vectors). *A matrix Y and its Cayley transform V commute. Hence they have the same eigenvectors. Their eigenvalues, denoted by $\{\mu_i\}$ and $\{\lambda_i\}$, obey*

$$\lambda_i = \frac{1-\mu_i}{1+\mu_i}. \tag{13}$$

Proof: Omitted. $\square$

Lemma 3 (Full Diversity). *A set of unitary matrices $\{V_0, \ldots, V_L\}$ is fully-diverse, i.e., $|\det(V_\ell - V_{\ell'})|$ is nonzero for all $\ell \neq \ell'$, if and only if the set of its skew-Hermitian Cayley transforms $\{Y_0, \ldots, Y_L\}$ is fully-diverse. Moreover, we have*

$$V_\ell - V_{\ell'} = 2(I+Y_\ell)^{-1}\left[Y_{\ell'} - Y_\ell\right](I+Y_{\ell'})^{-1}. \tag{14}$$

Proof: We need only prove (14). We have

$$\begin{aligned} V_\ell - V_{\ell'} &= (I+Y_\ell)^{-1}(I-Y_\ell) - (I-Y_{\ell'})(I+Y_{\ell'})^{-1} \\ &= (I+Y_\ell)^{-1}\left[(I-Y_\ell)(I+Y_{\ell'}) - (I+Y_\ell)(I-Y_{\ell'})\right](I+Y_{\ell'})^{-1} \\ &= 2(I+Y_\ell)^{-1}\left[Y_{\ell'} - Y_\ell\right](I+Y_{\ell'})^{-1}. \end{aligned}$$

$\square$

Thus, to design a fully-diverse set of unitary matrices we can design a fully-diverse set of skew-Hermitian matrices and then employ the Cayley transform. This design technique is used in an example in Section 3.

2.3 Cayley differential codes

Because the Cayley transform maps the nonlinear Stiefel manifold to the linear space of skew-Hermitian matrices (and vice-versa) it is convenient to encode data onto a skew-Hermitian matrix and then apply the Cayley transform to get a unitary matrix. It is most straightforward to encode the data linearly.

We call a *Cayley Differential (CD) code* one for which each unitary matrix is computed by the Cayley transform

$$V = (I + iA)^{-1}(I - iA),$$

where the Hermitian matrix A is given by

$$A = \sum_{q=1}^{Q} \alpha_q A_q, \tag{15}$$

where $\alpha_1, \ldots, \alpha_Q$ are real scalars (chosen from a set $\mathcal{A}$ with r possible values) and where A_q are *fixed* $M \times M$ complex Hermitian matrices.

The code is completely determined by the set of matrices $A_1, \ldots, A_Q$, which can be thought of as Hermitian basis matrices. Each individual codeword, on the other hand, is determined by our choice of the scalars $\alpha_1, \ldots, \alpha_Q$. Since each α_q may each take on r possible values, and the code occupies M channel uses, the transmission rate is $R = (Q/M)\log_2 r$. Finally, since an arbitrary $M \times M$ Hermitian matrix is parameterized by M^2 real variables, we have the constraint

$$Q \leq M^2. \tag{16}$$

In Section 2.4, as a consequence of our decoding algorithm, we shall impose a more stringent constraint on Q.

We defer discussion of how to choose Q and design the A_q's and the set $\mathcal{A}$ until Section 2.5. We concentrate now instead on how to decode $\alpha_1, \ldots, \alpha_Q$ at the receiver.

2.4 Decoding the CD codes

An important property of the CD codes is the ease with which the receiver may form a system of linear equations in the variables $\{\alpha_q\}$. To see this, it is useful to write the fundamental receiver equation (6) using the Cayley transform,

$$\begin{aligned} X_\tau &= V_{z_\tau} X_{\tau-1} + W_\tau - V_{z_\tau} W_{\tau-1} \\ &= (I+iA)^{-1}(I-iA)X_{\tau-1} + W_\tau - (I+iA)^{-1}(I-iA)W_{\tau-1} \end{aligned}$$

implying that

$$(I+iA)X_\tau = (I-iA)X_{\tau-1} + (I+iA)W_\tau - (I-iA)W_{\tau-1},$$

or

$$X_\tau - X_{\tau-1} = A\frac{1}{i}(X_\tau + X_{\tau-1}) + (I+iA)W_\tau - (I-iA)W_{\tau-1}, \tag{17}$$

which is linear in A. Since the data $\{\alpha_q\}$ is also linear in A, equation (17) is linear in $\{\alpha_q\}$.

We look first at maximum-likelihood estimation of the $\{\alpha_q\}$. Using (17) and noting that the additive noise $(I+iA)W_\tau - (I-iA)W_{\tau-1}$ has independent columns with covariance $2(I+iA)(I-iA) = 2(I+A^2)$ shows that the ML decoder is

$$\hat{\alpha}_{\mathrm{ml}} = \arg\min_{\{\alpha_q\}} \left\| (I+iA)^{-1}\left(X_\tau - X_{\tau-1} - \frac{1}{i}A(X_\tau + X_{\tau-1})\right)\right\|^2,$$

or, more explicitly,

$$\hat{\alpha}_{\mathrm{ml}} = \arg\min_{\{\alpha_q\}} \left\| \left(I + i\sum_{q=1}^{Q}\alpha_q A_q\right)^{-1} \left(X_\tau - X_{\tau-1} - \frac{1}{i}\sum_{q=1}^{Q}\alpha_q A_q (X_\tau + X_{\tau-1})\right)\right\|^2. \tag{18}$$

This decoder is not quadratic in $\{\alpha_q\}$ and so may be difficult to solve. However, if we ignore the covariance of the additive noise in (17) and assume that the noise is simply spatially white, then we obtain the linearized ML decoder

$$\hat{\alpha}_{\mathrm{lin}} = \arg\min_{\{\alpha_q\}} \left\| \left(X_\tau - X_{\tau-1} - \frac{1}{i}\sum_{q=1}^{Q}\alpha_q A_q (X_\tau + X_{\tau-1})\right)\right\|^2. \tag{19}$$

We call the decoder "linearized" because the system of equations obtained in solving (19) for unconstrained $\{\alpha_q\}$ is linear.

Because (19) is quadratic in $\{\alpha_q\}$, a simple approximate solution for $\{\alpha_q\}$ chosen from a fixed constellation can use nulling and cancelling (as in BLAST—see [24, 25, 26]). An exact solution without an exhaustive search can use sphere decoding [27, 28]. To facilitate the presentation of these decoding methods, we introduce some matrix notation.

2.4.1 An equivalent-channel model

Define $C = X_\tau - X_{\tau-1}$ and $B = -i(X_\tau + X_{\tau-1})$, and rewrite (17) as

$$C = AB + W_A = \sum_{q=1}^{Q} \alpha_q A_q B + W_A, \tag{20}$$

where $W_A = (I + iA)W_\tau - (I - iA)W_{\tau-1}$ is additive Gaussian noise where each column is independent and has mean zero and covariance $2(I + A^2)$. Equation (20) may be written in a more convenient form by decomposing the matrices into their real and imaginary parts to obtain

$$C_R + iC_I = \sum_{q=1}^{Q} \alpha_q (A_{R,q} + iA_{I,q})(B_R + iB_I) + W_{R,A} + iW_{I,A}$$

where

$$\begin{aligned} C_R &= \sum_{q=1}^{Q} (A_{R,q}B_R - A_{I,q}B_I)\alpha_q + W_{R,A} \\ C_I &= \sum_{q=1}^{Q} (A_{I,q}B_R + A_{R,q}B_I)\alpha_q + W_{I,A}. \end{aligned}$$

Denoting the columns of C_R, C_I, B_R, B_I, $W_{R,A}$, and $W_{I,A}$ by $c_{R,n}$, $c_{I,n}$, $b_{R,n}$, $b_{I,n}$, $w_{R,n,A}$, and $w_{I,n,A}$, where $n = 1, \ldots, N$, we gather the two above equations to form the single real system of equations

$$\underbrace{\begin{bmatrix} c_{R,1} \\ c_{I,1} \\ \vdots \\ c_{R,N} \\ c_{I,N} \end{bmatrix}}_{c} = \mathcal{B} \underbrace{\begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_Q \end{bmatrix}}_{\alpha} + \underbrace{\begin{bmatrix} w_{R,1,A} \\ w_{I,1,A} \\ \vdots \\ w_{R,N,A} \\ w_{I,N,A} \end{bmatrix}}_{w}, \tag{21}$$

where the $2MN \times Q$ real matrix $\mathcal{B}$ is

$$\mathcal{B} = \begin{bmatrix} \begin{bmatrix} A_{R,1} & -A_{I,1} \\ A_{I,1} & A_{R,1} \end{bmatrix} \begin{bmatrix} b_{R,1} \\ b_{I,1} \end{bmatrix} & \cdots & \begin{bmatrix} A_{R,Q} & -A_{I,Q} \\ A_{I,Q} & A_{R,Q} \end{bmatrix} \begin{bmatrix} b_{R,1} \\ b_{I,1} \end{bmatrix} \\ \vdots & \ddots & \vdots \\ \begin{bmatrix} A_{R,1} & -A_{I,1} \\ A_{I,1} & A_{R,1} \end{bmatrix} \begin{bmatrix} b_{R,N} \\ b_{I,N} \end{bmatrix} & \cdots & \begin{bmatrix} A_{R,Q} & -A_{I,Q} \\ A_{I,Q} & A_{R,Q} \end{bmatrix} \begin{bmatrix} b_{R,N} \\ b_{I,N} \end{bmatrix} \end{bmatrix}. \tag{22}$$

We have a linear relation between the input and output vectors α and c

$$c = \mathcal{B}\alpha + w, \tag{23}$$

where α appears to pass through an equivalent channel $\mathcal{B}$ that is *known* to the receiver because $\mathcal{B}$ is a function of $A_1, \ldots, A_Q$, X_τ, and $X_{\tau-1}$. (The receiver simply uses (22) to find the equivalent channel.)

If we ignore the dependence of the noise covariance on the signal A, which is equivalent to considering the linearized maximum-likelihood criterion (19), we have a simple linear system of equations that may be decoded using known techniques such as successive nulling and cancelling [24], its efficient square-root implementation [26], or sphere decoding [27, 28]. Efficient implementations of nulling and cancelling generally require $O(Q^3)$ computations. Sphere decoding can be regarded as a generalization of nulling and cancelling where at each step, rather than making a hard decision on the corresponding α_q, one considers all α_q that lie within a sphere of a certain radius. Sphere decoding has the important advantage over nulling and cancelling that it computes the *exact* solution to (19). It can be computationally more intense—its worst-case behavior is exponential in M—but its *average* behavior is comparable to nulling/cancelling. This is especially true at high SNR [28, 36]; our simulations in Section 3 that use sphere decoding show that the SNR generally need only be moderate. Our simulations also give some comparisons with nulling/cancelling. We have found that, on average, sphere decoding solves (19) in time that is polynomial in Q (or M).

When the number of transmit antennas is small (say $M \leq 4$) ML decoding is possible if the data rates are not too high. However, exact ML decoding, as given by (18), generally requires a search over all $\alpha_1, \ldots, \alpha_Q \in \mathcal{A}$, which may be impractical. Fortunately, as shown in Section 3, the performance penalty for linearizing the likelihood is small, especially when weighed against the complexity of exact ML. Finally, we mention that the solution of the linearized ML criterion can be used as an initial condition for Newton-Raphson-type methods applied to the true ML criterion.

2.4.2 The number of independent equations

Nulling and cancelling explicitly requires that the number of equations be at least as large as the number of unknowns. Sphere decoding does not have this hard constraint, but it benefits from more equations because the computational complexity grows exponentially in the difference between the number of unknowns and equations.[1] From equation (23), the matrix $\mathcal{B}$ has size $2MN \times Q$ and we therefore have $2MN$ equations and Q unknowns. Hence, we may impose the constraint

$$Q \leq 2MN, \tag{24}$$

This argument assumes that the matrix $\mathcal{B}$ has full column rank. There is, at first glance, no reason to assume otherwise but it turns out to be false.

Theorem 1 (Rank of $\mathcal{B}$). *The matrix $\mathcal{B}$ given in (22) generically has rank*

$$\operatorname{rank} \mathcal{B} = \begin{cases} \min(2MN - N^2, Q) & M \geq N \\ \min(M^2, Q) & M < N \end{cases} \tag{25}$$

Proof: First, assume that $M \geq N$. The rank of $\mathcal{B}$ is the dimension of the range space of c in the equation $c = \mathcal{B}\alpha$ as α varies. Equivalently, the rank of $\mathcal{B}$ is the dimension of the range space of C in the equation $C = AB$ as A varies. Since C is an $M \times N$ matrix it would appear that the range space of C has $2MN$ real dimensions. This would be true if A were an arbitrary matrix, but A is constrained to vary only over the space of Hermitian matrices. We study the consequences of this constraint.

First note that

$$B^*C = B^*AB.$$

On the other hand, $C^* = B^*A^* = B^*A$, or

$$C^*B = B^*AB = B^*C. \tag{26}$$

Therefore, the $N \times N$ matrix C^*B is Hermitian. This enforces N^2 linear constraints on the entries of C (N constraints to make the diagonals of C^*B real and $2\frac{N(N-1)}{2}$ constraints to make the upper and lower

[1] If this difference is not very large sphere decoding is still feasible. In Section 3 we consider an example where the number of unknowns is 16 and the number of equations 12.

triangular entries conjugates of one another). Therefore the $2MN$ entries of C are not all free: at most $2MN - N^2$ of them are. On the other hand, since α has Q entries (equivalently, A has Q degrees of freedom), the rank of $\mathcal{B}$ is therefore no greater than

$$\min(2MN - N^2, Q).$$

Our argument so far has not relied on a specific set of basis matrices $A_1, \ldots, A_Q$. However, for a generic choice of basis matrices the rank of $\mathcal{B}$ is given by $\min(2MN - N^2, Q)$, which is the desired result.

Assume now that $M < N$. We know that the $N \times N$ matrix C^*B is Hermitian, but it no longer has full rank—it has rank $M < N$. In particular, the entries of the lower right $(N - M) \times (N - M)$ Hermitian submatrix of C^*B are uniquely determined from its other entries. Therefore the equation $C^*B = B^*C$ yields $N^2 - (N - M)^2 = 2MN - M^2$ constraints (we remove the Hermitian constraints arising from the lower right $(N - M) \times (N - M)$ submatrix of C^*B because they are redundant). Thus, there are at most $2MN - (2MN - M^2) = M^2$ degrees of freedom in C. Therefore, the rank of $\mathcal{B}$ is

$$\min(M^2, Q)$$

which is the desired result. □

Theorem 1 shows that even though there are $2MN$ equations in (23), not all of them are independent. To have at least as many equations as unknowns when $M \geq N$, we therefore impose the constraint

$$Q \leq 2MN - N^2, \qquad (M \geq N). \tag{27}$$

When $M < N$, only M^2 of the equations in (23) are independent so

$$Q \leq M^2, \qquad (M < N). \tag{28}$$

Inequalities (27) and (28) can be combined into the single inequality

$$Q \leq K(2M - K), \quad \text{where} \quad K = \min(M, N). \tag{29}$$

2.5 Design of the CD Codes

Although we have introduced the CD structure

$$A = \sum_{q=1}^{Q} \alpha_q A_q,$$

we have not yet specified Q, nor have we explained how to design the Hermitian basis matrices $A_1, \ldots, A_Q$ or choose the discrete set $\mathcal{A}$ from which the α_q are drawn. We now address these issues.

2.5.1 Choice of Q

To make the constellation as rich as possible we should make the number of degrees of freedom Q as large as possible. Therefore, as a general practice, we find it useful to take Q at its upper limit in (29),

$$Q = \min(N, M)\max(2M - N, M). \tag{30}$$

If sphere decoding is used we sometimes exceed this limit (yielding more unknowns than equations; see examples in Section 3), but we always obey $Q \leq M^2$.

We are left with how to design $A_1, \ldots, A_Q$ and how to choose the discrete set $\mathcal{A}$. If the rates being considered are reasonably small (for example, $R < 4$) then the criterion given in [14] of maximizing $|\det(V_\ell - V_{\ell'})|$ for all $\ell' \neq \ell$ is tractable. Recall that any constellation $V_1, \ldots, V_L$ for which this determinant is nonzero for all $\ell' \neq \ell$ is said to be fully diverse. Lemma 3 shows that a constellation of unitary matrices is fully diverse if and only if the corresponding Cayley-transformed constellation of skew-Hermitian matrices is fully-diverse. Since

$$A' - A = \sum_{q=1}^{Q} A_q(\alpha'_q - \alpha_q),$$

by considering α and α' that differ in only one coordinate q, we see that it is necessary (but not sufficient) for $A_1, \ldots, A_Q$ to be nonsingular. We show some examples of full diversity for small rates and small number of antennas in Section 3.

At high rates, however, we do not pursue the full-diversity criterion. The reasons are two-fold: first, the criterion becomes intractable because of the number of matrices involved; second, the performance of the constellation may not be governed so much by its worst-case pairwise $|\det(V_\ell - V_{\ell'})|$, but rather how well the matrices are distributed throughout the space of unitary matrices. One reason why group constellations do not perform very well at high rates is because they lack the required statistical structure of a good high rate constellation [19].

2.5.2 The mutual-information-maximizing distribution

In [6], code design for the known channel requires the design of so-called *dispersion matrices*, which play a role similar to $A_1, \ldots, A_Q$ in our problem. To ensure that the resulting constellation has the correct statistical structure, the dispersion matrices are chosen to maximize the mutual information between the input and output signals. It is shown that maximizing mutual information also has a beneficial effect on the average probability of error [6] at high rates. We seek a similar quality criterion here.

Unfortunately, we cannot adopt this strategy directly to design $A_1, \ldots, A_Q$ because, unlike the known channel case, we do not know how to compute the mutual information between the input $\alpha_1, \ldots, \alpha_Q$ and output pair $(X_{\tau-1}, X_\tau)$ for the differential scheme. We can, however, approximate this strategy by choosing $A_1, \ldots, A_Q$ such that the distribution on V is close to the distribution that maximizes the mutual information between it and the pair $(X_{\tau-1}, X_\tau)$. We give the maximizing distribution for V in the following theorem.

Theorem 2 (Optimal distribution on V). *The mutual information between the unitary input matrix V and the output $(X_{\tau-1}, X_\tau)$ in the differential scheme*

$$\begin{bmatrix} X_{\tau-1} \\ X_\tau \end{bmatrix} = \sqrt{\rho} \begin{bmatrix} S_{\tau-1} \\ V S_{\tau-1} \end{bmatrix} H + \begin{bmatrix} W_{\tau-1} \\ W_\tau \end{bmatrix}, \tag{31}$$

where $S_{\tau-1} = V_{z_{\tau-1}} V_{z_{\tau-2}} \ldots$, and where H, $W_{\tau-1}$, and W_τ are $M \times N$ matrices with independent $\mathcal{CN}(0,1)$ entries, is maximized when V is isotropically distributed.

Proof: See Appendix A.

Remark: An *isotropically-distributed* (i.d.) unitary matrix V is one whose probability density function is invariant to pre- and post-multiplication by an arbitrary unitary matrix. That is,

$$p(V) = p(\Theta V) = p(V\Theta),$$

for all unitary Θ (see, e.g., [9, 12]). The probability density function of an isotropically-distributed unitary matrix is often referred to as the *Haar measure* or the uniform measure on the unitary group.

Hence, good constellations of unitary matrices have the appearance of being taken independently from an isotropic distribution.

2.5.3 Cauchy random matrices

Since our data modulates the A matrix, we would like to know the optimal distribution on A. Equivalently, we need to find the distribution on A that yields a $V = (I + iA)^{-1}(I - iA)$ that is isotropically distributed.

Theorem 3 (Optimal distribution on A). *The unitary matrix $V = (I + iA)^{-1}(I - iA)$ is isotropically-distributed if and only if the Hermitian matrix A has the distribution*

$$p(A) = \frac{2^{M^2-M}(M-1)!\cdots 1!}{\pi^{M(M+1)/2}} \frac{1}{\det(I + A^2)^M}. \tag{32}$$

Proof: See Appendix B.

The probability density function (32) is the matrix generalization of the familiar scalar Cauchy distribution

$$p(a) = \frac{1}{\pi(1 + a^2)}. \tag{33}$$

A scalar isotropic v can be written as $v = e^{i\theta}$, where θ is uniform over $[0, 2\pi)$. In this case, $a = -i\frac{1-e^{i\theta}}{1+e^{i\theta}} = -\tan(\theta/2)$ is Cauchy. The scalar Cauchy random variable is often regarded as the ratio of two independent Gaussian random variables. It is (in)famous because it has infinite variance, and the mean of n independent Cauchy random variables has the same Cauchy distribution—the law of large numbers does not apply. We refer to any random Hermitian matrix whose probability density function is (32) as a *Cauchy random matrix*.

2.5.4 Choice of $\mathcal{A}$

Theorem 3 implies that, at high rates, our CD code constellation $\mathcal{A} = \sum_{q=1}^{Q} A_q \alpha_q$ should resemble samples from a Cauchy random matrix distribution. We look first at the implications when there is one transmit antenna $M = 1$. In this case the optimal strategy is standard DPSK.

Theorem 3 when $M = 1$ gives us the scalar Cauchy density (33). By (29) we are limited to $Q = 1$, and there is no loss of generality in setting $A_1 = 1$ to get

$$v = \frac{1 - i\alpha_1}{1 + i\alpha_1}, \qquad \alpha_1 = -i\frac{1 - v}{1 + v}. \tag{34}$$

To get rate $R = (Q/M)\log_2 r$ with $M = Q = 1$ we need $\mathcal{A}$ to have $r = 2^R$ points. Standard DPSK puts these points uniformly around the unit circle at angular intervals of $2\pi/r$ with the first point at angle π/r. (The location of the first point does not affect the constellation performance in any way, but it helps us avoid a formal singularity in the inversion formula (34) at $v = -1$.) For a point at angle θ on the unit circle,

$$\alpha_1 = -i\frac{1-e^{i\theta}}{1+e^{i\theta}} = -\tan(\theta/2). \tag{35}$$

For example, for $r = 2$ (D-BPSK), we have $\mathcal{V} = \{e^{\pi i/2}, e^{-\pi i/2}\}$. Plugging these values into (35) yields $\mathcal{A} = \{-1, 1\}$. For $r = 4$, (D-QPSK), we have $\mathcal{A} = \{-1-\sqrt{2}, 1-\sqrt{2}, -1+\sqrt{2}, 1+\sqrt{2}\} = \{-2.4142, -0.4142, 0.4142, 2.4142\}$ (we always arrange the points in increasing order). For $r = 8$, $\mathcal{A} = \{-5.0273, -1.4966, -0.6682, -0.1989, 0.1989, 0.6682, 1.4966, 5.0273\}$. We see that the points rapidly spread themselves out as r increases, thus reflecting the long tail of the Cauchy distribution (33).

We denote $\mathcal{A}_r$ to be the image of the function (35) applied to the set $\theta \in \{\pi/r, 3\pi/r, 5\pi/r, \ldots, (2r-1)\pi/r\}$. In the limit as $r \to \infty$, the fraction of points in $\mathcal{A}_r$ less than some x is given by the cumulative Cauchy distribution evaluated at x. The set $\mathcal{A}_r$ can thus be regarded as an r-point discretization of a scalar Cauchy random variable.

While this argument tells us how to choose the set $\mathcal{A}$ as a function of r when $Q = M = 1$, it does not directly show us how to choose $\mathcal{A}$ when $M > 1$. Nevertheless, when $M > 1$ we also set $\mathcal{A} = \mathcal{A}_r$. Thus, the $\{\alpha_q\}$ are chosen as discretized scalar Cauchy random variables for any Q and M. To complete the code construction, it is crucial that $\{A_1, \ldots, A_Q\}$ be chosen appropriately, and we present a criterion in the next section.

2.5.5 Choice of $\{A_q\}$

We shift our attention away from the final distribution on A and express our design criterion in terms of V. For a given $A_1, \ldots, A_Q$ and $\mathcal{A}_r$, we define a distance criterion for the resulting constellation of matrices $\mathcal{V}$ to be

$$\xi(\mathcal{V}) = \frac{1}{M}\mathrm{E}\,\log\det(V-V')(V-V')^* = \frac{2}{M}\mathrm{E}\,\log|\det(V-V')|, \tag{36}$$

where $V' = (I+iA')^{-1}(I-iA')$, $A' = \sum_{q=1}^{Q} A_q\alpha'_q$, and the expectation is over $\alpha_1, \ldots, \alpha_Q$ and $\alpha'_1, \ldots, \alpha'_Q$ chosen uniformly from $\mathcal{A}_r$ such that $\alpha \neq \alpha'$. Although $\xi(\mathcal{V})$ is often negative, it is a measure of the expected "distance" between the random matrices V and V' and clearly is similar to the $|\det(V-V')|$ criterion in (1). If we interchange the expectation and the $\log(\cdot)$, the criterion directly measures the expected pairwise probability of error (9). Thus, maximizing $\xi(\mathcal{V})$ is connected with lowering average pairwise error probability. To choose the A_q's, we therefore propose the optimization problem

$$\arg \max_{A_q=A_q^*, q=1,\ldots,Q} \xi(\mathcal{V}). \tag{37}$$

Our choices of A_q and $\mathcal{A}_r$ affect the distance criterion through the distribution $p_V(\cdot)$ that they impose on the V matrices. To connect the optimization of this criterion with the information-theoretic considerations of Section 2.5.2, we prove the next theorem, which shows that this criterion is maximized when V and V' are independently chosen isotropic matrices.

Theorem 4 (Isotropic distribution maximizes criterion). *Let V and V' be $M \times M$ random unitary matrices chosen independently according to some common distribution $p_V(\cdot)$. Then the distance criterion (36) obeys*

$$\frac{1}{M}\mathrm{E}\,\log\det(V-V')(V-V')^* \leq 0, \tag{38}$$

with equality when $p_V(\cdot)$ is the isotropic distribution.

Proof: See Appendix D.

We interpret (36) as a measure of the average distance between matrices in the constellation. Theorem (4) says that if the set $\mathcal{A}_r$ and $A_1, \ldots, A_Q$ are chosen such that V is approximately isotropically distributed when $\mathcal{A}_r$ is sampled uniformly, then the average distance should be large.

We use (14), and the fact that matrices commute inside the determinant function, to write the optimization as a function of A and A',

$$\arg \max_{A_q=A_q^*, q=1,\ldots,Q} \left[\log 4 - \frac{1}{M}\mathrm{E}\,\log\det(I+A^2) - \frac{1}{M}\mathrm{E}\,\log\det(I+A'^2) + \frac{1}{M}\mathrm{E}\,\log\det(A-A')^2\right], \tag{39}$$

where $A = \sum_{q=1}^{Q} A_q\alpha_q$, $A' = \sum_{q=1}^{Q} A_q\alpha'_q$. For a constellation with $\alpha_1, \ldots, \alpha_Q$ and $\alpha'_1, \ldots, \alpha'_Q$ chosen from $\mathcal{A}_r$, we interpret the expectation as uniform over $\mathcal{A}_r$ such that $\alpha \neq \alpha'$.

It is occasionally useful, especially when r is large, to replace the discrete set from which α_q and α'_q are chosen ($\mathcal{A}_r$) with independent scalar Cauchy distributions. In this case, since the sum of two independent Cauchy random variables is scaled-Cauchy, our criterion simplifies to

$$\arg \max_{A_q=A_q^*, q=1,\ldots,Q} \left[2\log 4 - \frac{2}{M}\mathrm{E}\,\log\det(I+A^2) + \frac{1}{M}\mathrm{E}\,\log\det A^2\right], \tag{40}$$

where $A = \sum_{q=1}^{Q} A_q \alpha_q$ and the expectation is over $\alpha_1, \ldots, \alpha_Q$ chosen independently from a Cauchy distribution.

2.5.6 Design Method Summary

We now summarize the design method for a CD code with M transmit and N receive antennas, and target rate R.

(i) Choose $Q \leq \min(N, M) \max(2M - N, M)$. This inequality is a hard limit for decoding by nulling/cancelling, and Q is typically chosen to make it an equality. But the inequality is a soft limit for sphere decoding and we may choose Q as large as M^2 even if $N < M$.

(ii) Since $R = \frac{Q}{M} \log_2 r$, set $r = 2^{MR/Q}$. Let $\mathcal{A}_r$ be the r-point discretization of the scalar Cauchy distribution obtained as the image of the function (35) applied to the set $\theta \in \{\frac{\pi}{r}, \frac{3\pi}{r}, \ldots, \frac{(2r-1)\pi}{r}\}$.

(iii) Choose a set $\{A_q\}$ that solves the optimization problem (39).

We now make some remarks:

1. The solution to (39) is highly nonunique: simply reordering the $\{A_q\}$ gives another solution, as does changing the signs of the $\{A_q\}$, since the sets $\mathcal{A}_r$ are symmetric about the origin.

2. It does not appear that (39) has a simple closed-form solution for general Q, M, and N, but in Section 3 we give a special case where a closed-form solution appears.

3. We solve (39) numerically using gradient-ascent methods. The computation of the gradient of the criterion in (39) is presented in Appendix C. Since the criterion function is nonlinear and nonconcave in the design variables $\{A_q\}$, there is no guarantee of obtaining a global maximum. However, since the code design is performed off-line and only once, one can use more sophisticated optimization techniques that vary the initial condition, use second-order methods, use simulated annealing, etc. Section 3 shows that the codes obtained with a gradient search tend to have very good performance.

4. The entries of $\{A_q\}$ in (39) are unconstrained other than that the final matrix must be Hermitian. Appealing to symmetry arguments, however, we have found it beneficial to constrain the Frobenius norm of all the matrices in $\{A_q\}$ to be the same. In fact, in our experience it is very important both for the criterion function (39) and for the ultimate constellation performance that the correct Frobenius norm of the basis matrices be chosen. With the correct Frobenius norm, choosing an initial condition

for the $\{A_q\}$ in the gradient search becomes easier. The gradient for the Frobenius norm has a simple closed form which we now give. It can be used to solve for the optimal norm.

Let $\sqrt{\gamma}$ be a multiplicative factor that we use to multiply every A_q; we solve for the optimal $\gamma > 0$ by maximizing the criterion function (40) (the same analysis holds for (39))

$$\begin{aligned}
&\arg\max_{\gamma} \left[2\log 4 - \frac{2}{M}\mathrm{E}\ \log\det(I+\gamma A^2) + \frac{1}{M}\mathrm{E}\ \log\det\gamma A^2\right] \\
&\quad = \arg\max_{\gamma} \left[\log\gamma - \frac{2}{M}\mathrm{E}\ \log\det(I+\gamma A^2)\right]
\end{aligned}$$

The optimal γ therefore sets the gradient of this last equation to zero:

$$\begin{aligned}
0 &= \frac{1}{\gamma} - \frac{2}{M}\mathrm{E}\,\mathrm{tr}\,[(I+\gamma A^2)^{-1}A^2] \\
&= \frac{1}{\gamma} - \frac{2}{M}\mathrm{E}\,\mathrm{tr}\,\left[(I+\gamma A^2)^{-1})\frac{1}{\gamma}(I+\gamma A^2 - I)\right] \\
&= \frac{1}{\gamma}\left(1 - \frac{2}{M}\mathrm{E}\,\mathrm{tr}\,[I - (I+\gamma A^2)^{-1}]\right) \\
&= \frac{1}{\gamma}\left(-1 + \frac{2}{M}\mathrm{E}\,\mathrm{tr}\,(I+\gamma A^2)^{-1}\right).
\end{aligned}$$

The equation $-1 + \frac{2}{M}\mathrm{E}\,\mathrm{tr}\,(I+\gamma A^2)^{-1} = 0$ can readily be solved numerically.

5. The ultimate rate of the code depends on the number of signals sent Q, and the size of the constellation $\mathcal{A}_r$ from which $\alpha_1, \ldots, \alpha_Q$ are chosen. The code rate in bits/channel is

$$R = \frac{Q}{M}\log_2 r. \tag{41}$$

We generally choose r to be a power of two.

6. The design criterion (39) *depends explicitly on the number of receive antennas* N through the choice of Q. Hence, the optimal codes, for a given M, are different for different N.

7. The variable Q is essentially also a design variable. In our experience, the CD code performance is generally best when Q is chosen as large as possible. For example, a code with a given Q and r is likely to perform better than another code of the same rate that is obtained by halving Q and squaring r. Nevertheless, it is sometimes advantageous to choose a small Q to design a code of a specific rate.

8. If r is chosen a power of two, a standard gray-code assignment of bits to the symbols of the set $\mathcal{A}_r$ may be used.

9. The dispersion matrices $\{A_q\}$ are Hermitian and, in general, complex.

3 Examples of CD Codes and Performance

In this section, we simulate the performance of CD codes for various numbers of antennas and rates. The channel is modeled as quasi-static, where the fading matrix between the transmitter and receiver is constant (but unknown) between two successive channel uses. Two error events of interest include block errors, which correspond to errors in decoding the $M \times M$ matrices $V_1, \ldots, V_L$, and bit errors, which correspond to errors in decoding $\alpha_1, \ldots, \alpha_Q$. The bits are mapped to α_q with a gray code (see Section 2.5) and therefore a block error may correspond to only a few bit errors. In some examples, we compare the performance of linearized likelihood (sphere decoding) with true maximum likelihood and nulling/cancelling.

Simple example: $M = 2$**,** $R = 1$

For $M = 2$ transmit antennas and rate $R = 1$ the constellation has $L = 4$ elements. In this case, it turns out that no constellation can have ζ defined in (1) larger than $\zeta = \sqrt{2/3} = 0.8165$; see Prop. 3 in [37] for a proof of this result. The optimal constellation corresponds to a tetrahedron whose corners lie on the surface of a three-dimensional unit sphere, and one representation of it is given by the four unitary matrices

$$V_1 = \begin{bmatrix} \sqrt{\frac{1}{3}} + j\sqrt{\frac{2}{3}} & 0 \\ 0 & \sqrt{\frac{1}{3}} - j\sqrt{\frac{2}{3}} \end{bmatrix}, \; V_2 = \begin{bmatrix} -\sqrt{\frac{1}{3}} & \sqrt{\frac{2}{3}} \\ -\sqrt{\frac{2}{3}} & -\sqrt{\frac{1}{3}} \end{bmatrix},$$

$$V_3 = \begin{bmatrix} -\sqrt{\frac{1}{3}} & -\sqrt{\frac{2}{3}} \\ \sqrt{\frac{2}{3}} & -\sqrt{\frac{1}{3}} \end{bmatrix}, \; V_4 = \begin{bmatrix} \sqrt{\frac{1}{3}} - j\sqrt{\frac{2}{3}} & 0 \\ 0 & \sqrt{\frac{1}{3}} + j\sqrt{\frac{2}{3}} \end{bmatrix}. \tag{42}$$

There are many equivalent representations, but it turns out that this particular choice can be constructed as a CD code with $Q = r = 2$, and the basis matrices are

$$A_1 = \begin{bmatrix} \frac{1}{\sqrt{2}(\sqrt{3}+1)} & \frac{-i}{\sqrt{2}(\sqrt{3}-1)} \\ \frac{i}{\sqrt{2}(\sqrt{3}-1)} & \frac{-1}{\sqrt{2}(\sqrt{3}+1)} \end{bmatrix} \quad \text{and} \quad A_2 = \begin{bmatrix} \frac{1}{\sqrt{2}(\sqrt{3}+1)} & \frac{i}{\sqrt{2}(\sqrt{3}-1)} \\ \frac{-i}{\sqrt{2}(\sqrt{3}-1)} & \frac{-1}{\sqrt{2}(\sqrt{3}+1)} \end{bmatrix}. \tag{43}$$

The matrices (42) are generated as the Cayley transform (11) of $A = A_1\alpha_1 + A_2\alpha_2$, with $\alpha_1, \alpha_2 \in \mathcal{A}_2 = \{-1, 1\}$.

For comparison, we may consider the constellation based on orthogonal designs for $M = 2$ and $R = 1$ used in [16] given by

$$V_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}, \quad V_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} -1 & 1 \\ -1 & -1 \end{bmatrix}, \quad V_3 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}, \quad V_4 = \frac{1}{\sqrt{2}}\begin{bmatrix} -1 & -1 \\ 1 & -1 \end{bmatrix} \tag{44}$$

which has $\zeta = 1/\sqrt{2} \approx 0.7071$, or the constellation given in [14]

$$V_\ell = \frac{1}{\sqrt{2}}\begin{bmatrix} e^{2\pi i/4} & 0 \\ 0 & e^{2\pi i/4} \end{bmatrix}^{\ell-1}, \quad \ell = 1, \ldots, 4$$

which also has $\zeta = 0.7071$. Since we are more interested in high rate examples, we do not plot the performance of the CD code (42); however, simulations show that the performance gain over (44) is approximately 0.75 dB at high SNR. This small example shows that there are good codes within the CD structure at low rates. (In this case, the best $R = 1$ code has a CD structure.)

CD Code using orthogonal designs: $M = 2$

Recall from Lemma 3 that a constellation of unitary matrices is fully-diverse if and only if its Cayley transform constellation of skew-Hermitian matrices is fully-diverse. For $M = 2$, a famous fully-diverse constellation is the orthogonal design of Alamouti [4]

$$\mathrm{OD} = \begin{bmatrix} x & y \\ -y^* & x^* \end{bmatrix}. \tag{45}$$

Orthogonal designs are readily seen to be fully-diverse since

$$\det(\mathrm{OD}_1 - \mathrm{OD}_2) = \det\begin{bmatrix} x_1 - x_2 & y_1 - y_2 \\ -(y_1 - y_2)^* & (x_1 - x_2)^* \end{bmatrix} = |x_1 - x_2|^2 + |y_1 - y_2|^2.$$

We require that OD be skew-Hermitian, implying that

$$\mathrm{OD} = \begin{bmatrix} i\alpha_1 & \alpha_2 + i\alpha_3 \\ -\alpha_2 + i\alpha_3 & -i\alpha_1 \end{bmatrix} = i \underbrace{\begin{bmatrix} \alpha_1 & -i\alpha_2 + \alpha_3 \\ i\alpha_2 + \alpha_3 & -\alpha_1 \end{bmatrix}}_{=A}, \tag{46}$$

where the α_q's are real. Thus, we may define a CD code with basis matrices

$$A_1 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \quad A_3 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{47}$$

that generates a fully-diverse constellation. (In passing, we mention that A_1, A_2, and A_3 form a basis for the real Lie algebra $su(2)$ of traceless Hermitian matrices.) Using (14) yields

$$\det(V - V') = 4\det(I + iA)^{-1}\det(A' - A)\det(I + A')^{-1},$$

which upon simplification yields

$$\det(V - V') = \frac{4(|\alpha_1 - \alpha_1'|^2 + |\alpha_2 - \alpha_2'|^2 + |\alpha_3 - \alpha_3'|^2)}{(1 + |\alpha_1|^2 + |\alpha_2|^2 + |\alpha_3|^2)(1 + |\alpha_1'|^2 + |\alpha_2'|^2 + |\alpha_3'|^2)}. \tag{48}$$

For example, by choosing $\alpha_q \in \mathcal{A}_2$, we get a code with rate $R = 1.5$. The appropriate scaling (see Remark 4 in Section 2.5.6) is $\gamma = 1/3$. The resulting constellation of eight matrices (which we omit) has $\zeta = 1/\sqrt{3}$.

We note that the code (47) also appears to be a closed-form solution to (39) for $M = 2$ and $Q = 3$ because it is a local maximum to the criterion.

CD Code vs. OD: $M = N = 2$

For a higher-rate example, we examine another code for $M = 2$, but we choose $N = 2$ and $R = 6$. Figure 1 shows the performance of a CD code with $Q = 4$. The code is

$$\begin{aligned} A_1 &= \begin{bmatrix} 0.1785 & 0.0510 + 0.1340i \\ 0.0510 - 0.1340i & 0.0321 \end{bmatrix}, \quad A_2 = \begin{bmatrix} -0.1902 & 0.1230 + 0.0495i \\ 0.1230 - 0.0495i & -0.0512 \end{bmatrix}, \\ A_3 &= \begin{bmatrix} -0.2350 & 0.0515 - 0.0139i \\ 0.0515 + 0.0139i & 0.1142 \end{bmatrix}, \quad A_4 = \begin{bmatrix} 0.0208 & 0.1143 - 0.1532i \\ 0.1143 + 0.1532i & 0.0220 \end{bmatrix}. \end{aligned}$$

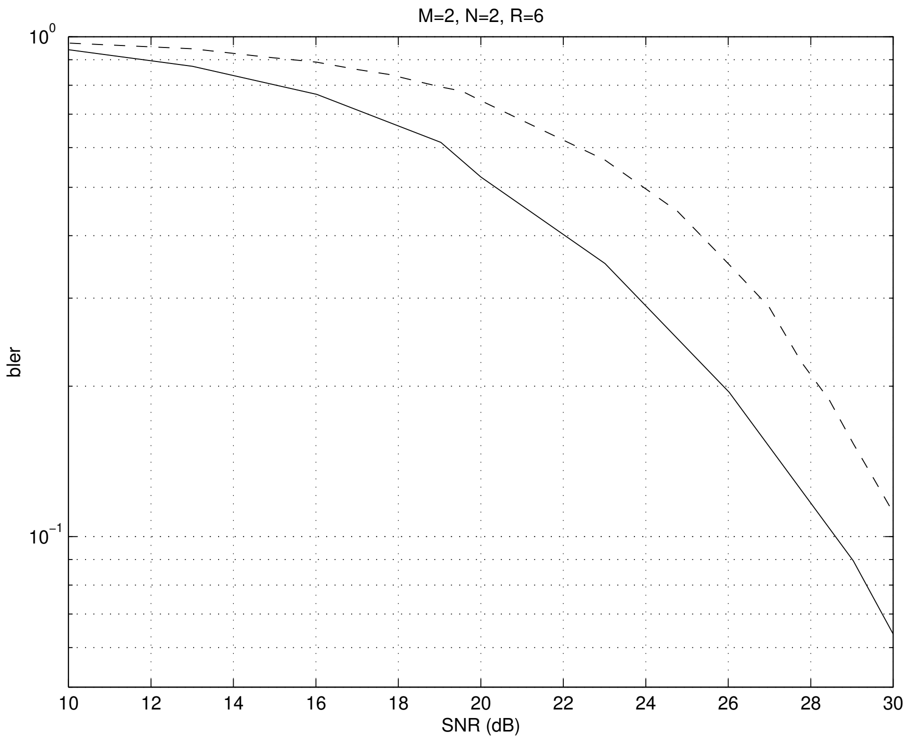

Figure 1: Performance of a CD code for $M = 2$ transmit and $N = 2$ receive antennas with rate $R = 6$. The solid line is the block error rate for the CD code with sphere decoding, and the dashed line is the differential two-antenna orthogonal design with maximum likelihood decoding.

(To get $R = 6$, choose $\alpha_1, \ldots, \alpha_4 \in \mathcal{A}_8$; the distance criterion (36) for this code is $\xi = -1.46$.) Also included in the figure is the two-antenna differential orthogonal design [16] with the same rate. The CD code obeys the constraint (29) and therefore can be decoded very quickly using the sphere decoder. A maximum likelihood decoder would have to search over $2^{RM} = 2^{12} = 4096$ matrices.

Comparison with another nongroup code: $M = 4$**,** $N = 1$**,** $R = 4$

There are not many performance curves easily available for existing codes for $M = R = 4$ over an unknown channel, but [19] has a nongroup code for $N = 1$ that appears in Table 4 and Figure 9 of that paper. Figure 2 compares it to a CD code with the same parameters. The CD code has $Q = 16$, and achieves $R = 4$ by choosing $r = 2$. The 4×4 matrices $A_1, \ldots, A_{16}$ are not given here, but they are available from the authors on demand; $\xi(\mathcal{V}) = -1.45$. The nongroup code, which has its origin in a group code, performs better but the difference is very small. Observe that $Q = M^2 > 2MN - N^2 = 7$ and therefore the inequality (29)

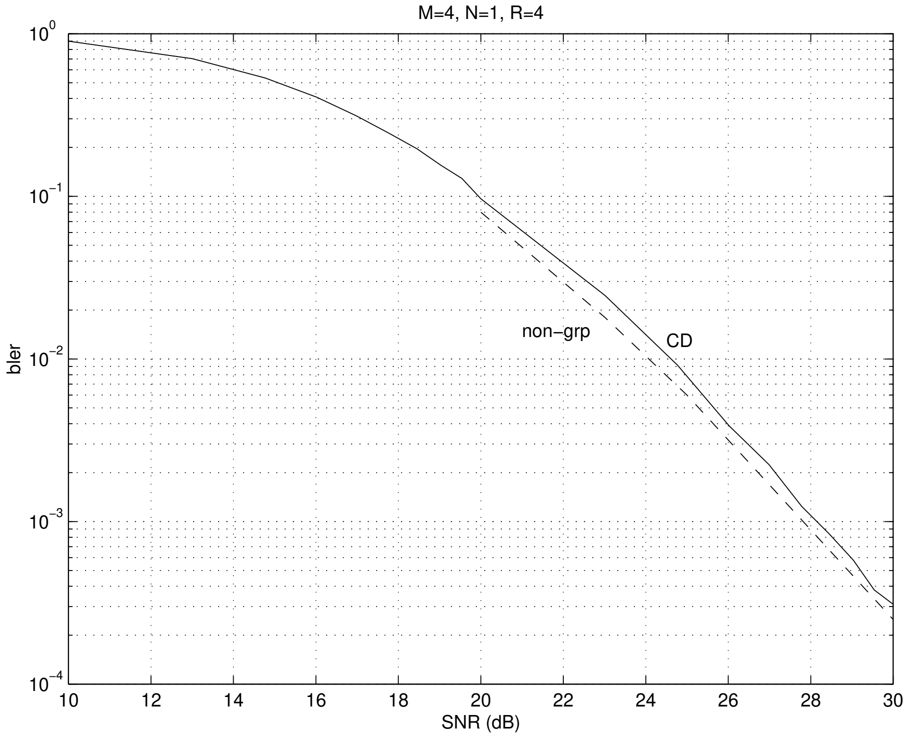

Figure 2: Block error performance of a CD code for $M = 4$ transmit and $N = 1$ receive antennas with rate $R = 4$ compared with nongroup code presented in [19] (Figure 9, Table 4). The decoding in both cases is true maximum likelihood through exhaustive search.

is not satisfied, but it does not matter in this case because the decoding for both codes is true maximum likelihood (rather than sphere decoding or nulling/cancelling). This example is not very practical because maximum likelihood decoding involves a search over $2^{RM} = 2^{16} = 65,536$ matrices. However, this same CD code is used in the next example where by increasing the number of receive antennas to $N = 2$ we are able to solve the linearized likelihood with sphere decoding.

Linearized vs. exact ML: $M = 4$**,** $N = 2$**,** $R = 4$

By increasing the number of receive antennas in the previous example to $N = 2$, we may linearize the likelihood and compare the performance with the true maximum likelihood. Figure 3 shows the results. (We use the same code as in Figure 2.) Observe that $Q > 2MN - N^2 = 12$ and therefore the inequality (29) is still not obeyed; but because it is almost obeyed, the sphere decoder of the linearized likelihood searches over only $16 - 12 = 4$ dimensions. With $r = 2$, this search is over $2^4 = 16$ quantities, which is

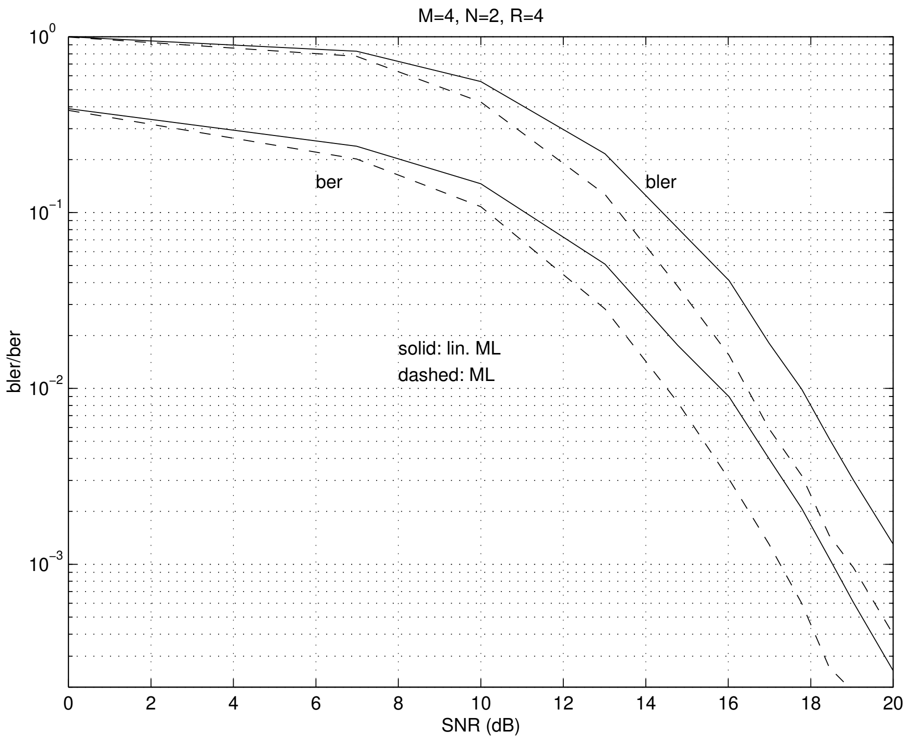

Figure 3: Performance of a CD code for $M = 4$ transmit and $N = 2$ receive antennas with rate $R = 4$. The solid lines are the block/bit error rates for the CD code with sphere decoding, and the dashed lines are the block/bit error rates with maximum likelihood decoding. The performance loss of linearizing the likelihood is approximately 1.3 dB at high SNR.

a negligible burden. Compare this burden with the true maximum likelihood (65,536 matrices). The figure shows that the performance loss for linearizing the likelihood is approximately 1.3 dB at high SNR. While the performance of linearized maximum likelihood is slightly worse than true maximum likelihood, the next figure shows that the performance of nulling/cancelling is much worse than either.

Sphere decoding vs. nulling/cancelling: $M = N = 4$, $R = 8$

Figure 4 shows the performance of a CD code for $M = 4$ transmit and $N = 4$ receive antennas for rate $R = 8$ with linearized-likelihood decoding. As in the previous example, $Q = 16$, but to achieve $R = 8$ we choose $r = 4$. (We again omit the explicit description of $A_1, \ldots, A_{16}$; $\xi(\mathcal{V}) = -1.36$.) Plotted also is a comparison of the same CD code with nulling/cancelling decoding. We see that sphere decoding is dramatically better. True maximum likelihood decoding is not realistic in this example because there are $2^{RM} = 2^{32} \approx 4 \times 10^9$ matrices in the codebook.

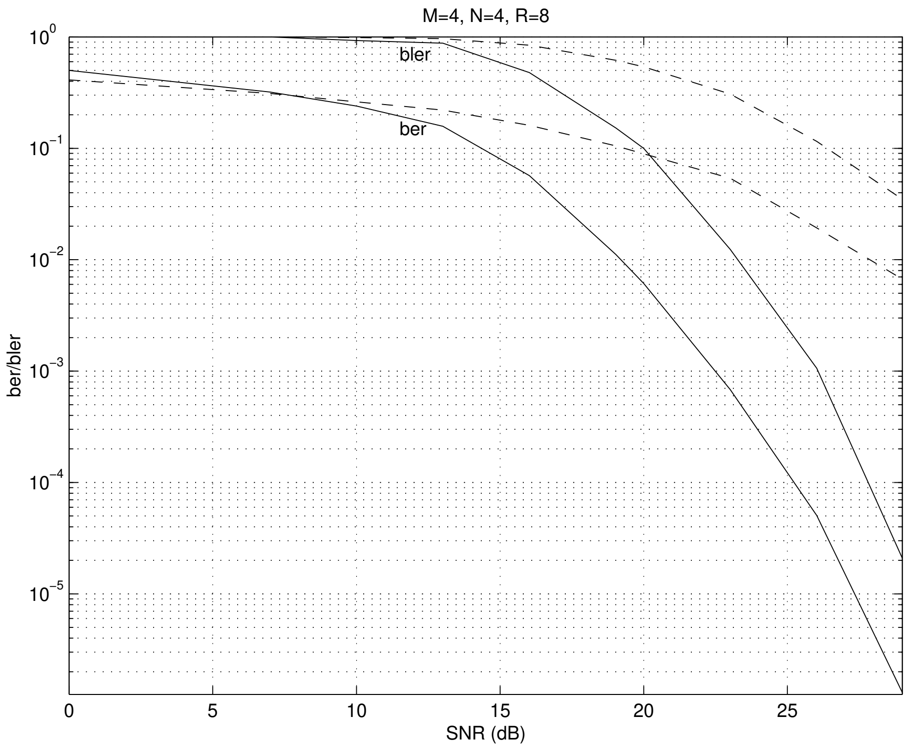

Figure 4: Performance of a CD code for $M = 4$ transmit and $N = 4$ receive antennas with rate $R = 8$. The solid lines are the block and bit error rates for sphere decoding and the dashed lines are for nulling/cancelling. The performance advantage of sphere decoding is dramatic.

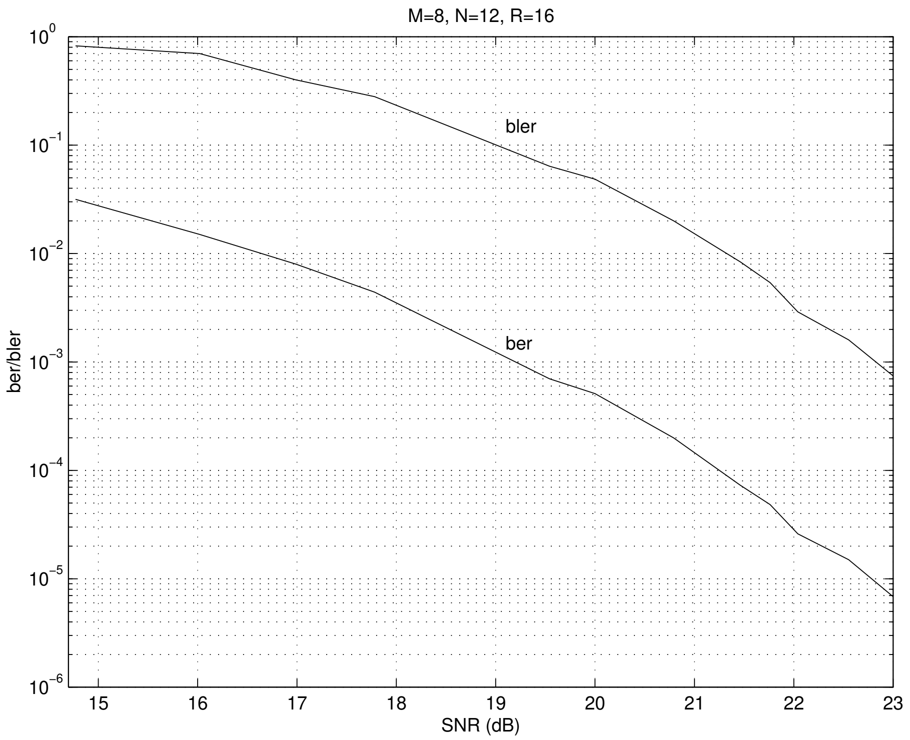

Figure 5: Performance of a CD code for $M = 8$ transmit and $N = 12$ receive antennas with rate $R = 16$. The solid lines are the block and bit error rates for sphere decoding.

High-rate example: $M = 8$**,** $N = 12$**,** $R = 16$

Some of the original V-BLAST experiments [24, 25] use eight transmit and twelve receive antennas to transmit more than 20 bits/second/Hz. Figure 5 shows that high rates with reasonable decoding complexity are also within reach of the CD codes. Plotted are the block and bit error rates for $R = 16$; here $Q = 64$ and $r = 4$ and the CD matrices are again omitted (they are available from the authors, and have $\xi(\mathcal{V}) = -1.48$). We note that because $M = 8$, the effective constellation size of unitary matrices is $L = 2^{RM} = 2^{128} \approx 3.4 \times 10^{38}$, yet we may still easily sphere decode the linearized likelihood.

4 Conclusion

The Cayley differential codes we have introduced do not require channel knowledge at the receiver, are simple to encode and decode, apply to any combination of transmit and receive antennas, and have excellent performance at very high rates. They are designed with a probabilistic criterion: they maximize the expected log-determinant of the difference between matrix pairs. The codes make use of the Cayley transform that maps the nonlinear Stiefel manifold of unitary matrices to the linear space of skew-Hermitian matrices. The transmitted data is broken into substreams $\alpha_1, \ldots, \alpha_Q$ and then linearly encoded in the Cayley transform domain. We showed that $\alpha_1, \ldots, \alpha_Q$ appear linearly at the receiver and can be decoded by nulling/cancelling or sphere decoding by ignoring the data dependence of the additive noise. Additional channel coding across $\alpha_1, \ldots, \alpha_Q$ or from block to block can be combined with a CD code to lower the error probability even further.

We have given some specific examples of the CD codes to indicate their performance, and presented a recipe for generating more codes for any combination of transmit and receive antennas. Our simulations indicate that codes generated with this recipe compare favorably with existing space-time schemes in their good performance and low decoding complexity.

In our simulations, we decoded the CD codes by maximum likelihood, sphere decoding of the linearized likelihood, or by nulling and cancelling. Sphere decoding of the linearized likelihood performed slightly worse than true maximum likelihood, but much better than nulling and cancelling and generally has comparable $O(Q^3)$ complexity. Exact maximum likelihood is generally not practical except with a small number of antennas or low rates. It may be possible to use the the sphere decoder output as the initial estimate for nonlinear second-order methods applied to the true ML criterion.

Our criterion function (37) was chosen for its ease of manipulation, and its connections to both minimizing error probability and achieving a constellation that is isotropically distributed. Nevertheless, although we generally found that high values of (37) led to good constellation performance, the criterion is not perfect—there were occasions where a larger value for (37) did not mean better performance. We may therefore ask if there are better criteria.

It would be interesting to see if the CD codes that satisfy the optimization (37) possess any general algebraic structure (Section 3 shows some cases where there is structure). This would lead to better theoretical understanding of the codes, as well as to possibly even faster and better decoding. There are potentially many ways to optimize (39), and the gradient method we chose is only one of them. More sophisticated optimization techniques may also be useful.

Our model assumed that the fading between the transmitter and receiver did not change between successive channel uses. In a more realistic model with a mobile transmitter or receiver, the channel would vary continuously from one use to the next. More analyses or simulations are needed to see how the performance would be affected by a varying channel—preliminary results suggest that the primary effect would be an "error floor" at very high SNR [38].

Finally, we chose the α_q's from a set $\mathcal{A}_r$ that is designed to help make the final A matrix behave, on average, like a Cauchy random matrix. We have not tried to optimize this set for code performance and think that this is another possible area for future work.

Acknowledgment

We are grateful to M. Oussama Damen for providing us with a preprint of [36], to J. Mazo for helping us prove Theorem 4, and to a reviewer's suggestion that lead to the first example in Section 3.

A Optimal input distribution

Let Θ and Φ^* be arbitrary fixed unitary matrices and write $V = \Theta V' \Phi^*$ for some V'. We rewrite equation (31) and substitute for V:

$$\begin{aligned} \begin{bmatrix} X_{\tau-1} \\ X_\tau \end{bmatrix} &= \sqrt{\rho} \begin{bmatrix} S_{\tau-1} \\ \Theta V' \Phi^* S_{\tau-1} \end{bmatrix} H + \begin{bmatrix} W_{\tau-1} \\ W_\tau \end{bmatrix} \\ &= \sqrt{\rho} \begin{bmatrix} \Phi\Phi^* S_{\tau-1} \\ \Theta V' \Phi^* S_{\tau-1} \end{bmatrix} H + \begin{bmatrix} W_{\tau-1} \\ W_\tau \end{bmatrix}. \end{aligned}$$

Pre-multiplying the first block equation by Φ^* and the second block equation by Θ^*, and noting that $S'_{\tau-1} = \Phi^* S_{\tau-1}$ is unitary and that $W'_{\tau-1} = \Phi^* W_{\tau-1}$ and $W'_\tau = \Theta^* W_\tau$ have the same distribution as $W_{\tau-1}$ and W_τ, we may write

$$\begin{bmatrix} \Phi^* X_{\tau-1} \\ \Theta^* X_\tau \end{bmatrix} = \sqrt{\rho} \begin{bmatrix} S'_{\tau-1} \\ V' S'_{\tau-1} \end{bmatrix} H + \begin{bmatrix} W'_{\tau-1} \\ W'_\tau \end{bmatrix}.$$

The joint distribution of H, $W'_{\tau-1}$ and W'_τ has not changed, and thus

$$p(X_{\tau-1}, X_\tau | V) = p(\Phi^* X_{\tau-1}, \Theta^* X_\tau | V') = p(\Phi^* X_{\tau-1}, \Theta^* X_\tau | \Theta^* V \Phi). \tag{A.1}$$

Assume that the mutual information $I(X_{\tau-1}, X_\tau; V)$ is maximized by some input distribution $p^*(V)$. For any fixed and unitary Θ and Φ, we have

$$\begin{aligned} I(X_{\tau-1}, X_\tau; V) &= I(\Phi^* X_{\tau-1}, \Theta^* X_\tau; V) \\ &= H(\Phi^* X_{\tau-1}, \Theta^* X_\tau) - H(\Phi^* X_{\tau-1}, \Theta^* X_\tau | V) \end{aligned}$$

$$
\begin{aligned}
&= H(X_{\tau-1}, X_\tau) - \int dX_{\tau-1} dX_\tau dV p^*(V) p(\Phi^* X_{\tau-1}, \Theta^* X_\tau | V) \log p(\Phi^* X_{\tau-1}, \Theta^* X_\tau | V) \\
&= H(X_{\tau-1}, X_\tau) - \int dX_{\tau-1} dX_\tau dV p^*(V) p(X_{\tau-1}, X_\tau | \Theta V \Phi^*) \log p(X_{\tau-1}, X_\tau | \Theta V \Phi^*) \\
&= H(X_{\tau-1}, X_\tau) - \int dX_{\tau-1} dX_\tau dV' p^*(\Theta^* V' \Phi) p(X_{\tau-1}, X_\tau | V') \log p(X_{\tau-1}, X_\tau | V')
\end{aligned}
$$

where in the fourth step we use (A.1), and in the last step the change of variables $V' = \Theta V \Phi$ (which has Jacobian determinant one). Hence, the input distribution $p^*(\Theta^* V \Phi)$ also maximizes the mutual information. The mutual information is concave as a function of the input distribution $p_V(\cdot)$. We conclude that the distribution $p^{**}(V) = K \int d\Theta d\Phi p^*(\Theta^* V \Phi)$, where the integral is over the space of unitary matrices Φ and Θ (i.e. over Haar measure, where K is the appropriate normalizing constant), also maximizes the mutual information. But $p^{**}(V)$ is clearly invariant to pre-and post-multiplication by fixed unitary matrices. Therefore, $p^{**}(V)$ is the isotropic distribution on the unitary matrix V.

B The Cauchy random matrix

We note from Lemma 2 that the Hermitian matrix A and the unitary matrix $V = (I + iA)^{-1}(I - iA)$ commute, and are therefore simultaneously diagonalized by a common set of orthonormal eigenvectors. We therefore first derive the distribution of the eigenvalues of A.

Let the eigenvalues of the $M \times M$ isotropically distributed unitary matrix V be denoted $\lambda = (\lambda_1, \ldots, \lambda_M)$, and let the eigenvalues of A be denoted $\mu = (\mu_1, \ldots, \mu_M)$. Then

$$
p(\lambda) = \frac{1}{M! \pi^M} \prod_{m=1}^{M} \delta(|\lambda_m|^2 - 1) \prod_{\ell > m} |\lambda_\ell - \lambda_m|^2. \tag{B.1}
$$

(See, for example, [30].) To get $p(\mu)$ we use the relations between the eigenvalues from Lemma 2

$$
\lambda_m = r_m \frac{1 - i\mu_m}{1 + i\mu_m}, \qquad d\lambda_m = r_m \frac{-2i}{(1 + i\mu_m)^2} d\mu_m,
$$

where $r_m = |\lambda_m|$. In general, μ is real and λ is complex, and we will integrate out $r = (r_1, \ldots, r_M)$. The Jacobian of the transformation $\lambda_m \mapsto (r_m, \mu_m)$ is $2r_m/(1 + \mu_m^2)$. The $\delta(\cdot)$ function in (B.1) becomes

$$
\delta(|\lambda_m|^2 - 1) = \delta(r_m^2 - 1) = \delta((r_m + 1)(r_m - 1)) = \frac{1}{2}\delta(r_m - 1),
$$

and we get

$$p(\mu, r) = \frac{1}{M!\pi^M} \prod_{m=1}^{M} \frac{r_m}{1+\mu_m^2} \delta(r_m - 1) \prod_{\ell > m} \left| r_\ell \frac{1 - i\mu_\ell}{1 + i\mu_\ell} - r_m \frac{1 - i\mu_m}{1 + i\mu_m} \right|^2.$$

Integrating out r gives the distribution on the eigenvalues

$$p(\mu) = \frac{1}{M!\pi^M} \prod_{m=1}^{M} \frac{1}{1+\mu_m^2} \prod_{\ell > m} \left| \frac{1 - i\mu_\ell}{1 + i\mu_\ell} - \frac{1 - i\mu_m}{1 + i\mu_m} \right|^2. \tag{B.2}$$

$$= \frac{1}{M!\pi^M} \prod_{m=1}^{M} \frac{1}{1+\mu_m^2} \prod_{\ell > m} \frac{4(\mu_\ell - \mu_m)^2}{(1+\mu_\ell^2)(1+\mu_m^2)} \tag{B.3}$$

$$= \frac{2^{M^2 - M}}{M!\pi^M} \prod_{m=1}^{M} \frac{1}{(1+\mu_m^2)^M} \prod_{\ell > m} (\mu_\ell - \mu_m)^2. \tag{B.4}$$

We now can obtain the distribution on A, by using results in [39, 40] that if a Hermitian matrix A has a distribution $p_A(A)$ that is invariant to unitary similarity transformations, then the joint density of the eigenvalues is

$$p(\mu) = \frac{\pi^{M(M-1)/2}}{M!(M-1)! \cdots 1} p_A(\text{diag}\,(\mu)) \prod_{\ell > m} (\mu_\ell - \mu_m)^2.$$

The distribution on A is invariant to unitary similarity transformations because

$$U^* A U = -iU^*(I+V)^{-1}(I-V)U = -iU^*(I+V)^{-1}UU^*(I-V)U = -i(I+U^*VU)^{-1}(I-U^*VU),$$

and the distribution of U^*VU is the same as the distribution of V. It follows that

$$p_A(A) = \frac{2^{M^2-M}(M-1)! \cdots 1!}{\pi^{M(M+1)/2}} \frac{1}{\det(I+A^2)^M}.$$

C Gradient of criterion (39)

In all the simulations presented in this paper the maximization of the design criterion function in (39), needed to design the CD codes, is performed using a simple constrained-gradient-ascent method. In this section, we compute the gradient of (39) that this method requires. More sophisticated optimization techniques that we do not consider, such as Newton-Raphson, scoring, and interior-point methods, can also use this gradient.

From (39), the criterion function is

$$\frac{1}{M} \mathrm{E} \log\det(V - V')(V - V')^* = \log 4 - \frac{1}{M} \mathrm{E} \log\det(I + A^2) - \frac{1}{M} \mathrm{E} \log\det(I + A'^2)$$

$$+\frac{1}{M}\mathrm{E}\,\log\det(A'-A)^2, \tag{C.1}$$

where $A = \sum_{q=1}^{Q} A_q\alpha_q$ and $A' = \sum_{q=1}^{Q} A_q\alpha'_q$. We are interested in the gradient of this function with respect to the matrices $A_1,\ldots,A_Q$. To compute the gradient of a real function $f(A_q)$ with respect to the entries of the Hermitian matrix A_q, we use the formulas

$$\left[\frac{\partial f(A_q)}{\partial \mathrm{Re}\, A_q}\right]_{j,k} = \lim_{\delta\to 0}\frac{1}{\delta}[f(A_q+\delta(e_je_k^T+e_ke_j^T))-f(A_q)], \qquad j\neq k \tag{C.2}$$

$$\left[\frac{\partial f(A_q)}{\partial \mathrm{Im}\, A_q}\right]_{j,k} = \lim_{\delta\to 0}\frac{1}{\delta}[f(A_q+i\delta(e_je_k^T-e_ke_j^T))-f(A_q)], \qquad j\neq k \tag{C.3}$$

$$\left[\frac{\partial f(A_q)}{\partial A_q}\right]_{j,j} = \lim_{\delta\to 0}\frac{1}{\delta}[f(A_q+\delta e_je_j^T)-f(A_q)], \tag{C.4}$$

where e_j is the M-dimensional unit column vector with a one in the jth entry and zeros elsewhere.

To apply (C.2) to the second term in (C.1), we compute

$$\begin{aligned}
&\log\det(I+(A+(e_je_k^T+e_ke_j^T)\alpha_q\delta)^2)\\
&\quad=\log\det(I+A^2+[A(e_je_k^T+e_ke_j^T)+(e_je_k^T+e_ke_j^T)A]\alpha_q\delta+O(\delta^2))\\
&\quad=\log\det\left[(I+A^2)(I+(I+A^2)^{-1}[A(e_je_k^T+e_ke_j^T)+(e_je_k^T+e_ke_j^T)A]\alpha_q\delta+O(\delta^2))\right]\\
&\quad=\log\det(I+A^2)+\mathrm{tr}\ \log(I+(I+A^2)^{-1}[A(e_je_k^T+e_ke_j^T)+(e_je_k^T+e_ke_j^T)A]\alpha_q\delta+O(\delta^2))\\
&\quad=\log\det(I+A^2)+\mathrm{tr}\ \left[(I+A^2)^{-1}[A(e_je_k^T+e_ke_j^T)+(e_je_k^T+e_ke_j^T)A]\alpha_q\delta\right]+O(\delta^2)\\
&\quad=\log\det(I+A^2)+\big[[(I+A^2)^{-1}A]_{k,j}+[(I+A^2)^{-1}A]_{j,k}+[A(I+A^2)^{-1}]_{k,j}\\
&\quad\ +[A(I+A^2)^{-1}]_{j,k}\big]\,\alpha_q\delta+O(\delta^2)\\
&\quad=\log\det(I+A^2)+4\mathrm{Re}\ \left[(I+A^2)^{-1}A\right]_{j,k}\alpha_q\delta+O(\delta^2).
\end{aligned}$$

The last equality follows because $(I+A^2)^{-1}$ and A commute and A is Hermitian. We may now apply (C.2) to obtain

$$\left[\frac{\partial \mathrm{E}\ \log\det(I+A^2)}{\partial \mathrm{Re}\, A_q}\right]_{j,k} = 4\mathrm{E}\,\mathrm{Re}\ \left[(I+A^2)^{-1}A\right]_{j,k}\alpha_q, \qquad j\neq k.$$

The gradient with respect to the imaginary components of A_q is handled in a similar way to obtain

$$\begin{aligned}
&\log\det(I+(A+(e_je_k^T-e_ke_j^T)\alpha_q i\delta)^2)\\
&\quad=\log\det(I+A^2)+\mathrm{tr}\ \left[(I+A^2)^{-1}[A(e_je_k^T-e_ke_j^T)+(e_je_k^T-e_ke_j^T)A]\alpha_q i\delta\right]+O(\delta^2)
\end{aligned}$$

$$
\begin{aligned}
&= \log\det(I + A^2) + \big[[(I + A^2)^{-1}A]_{k,j} - [(I + A^2)^{-1}A]_{j,k} + [A(I + A^2)^{-1}]_{k,j} \\
&\quad - [A(I + A^2)^{-1}]_{j,k}\big]\, \alpha_q i\delta + O(\delta^2) \\
&= \log\det(I + A^2) + 4\mathrm{Im}\, \left[(I + A^2)^{-1}A\right]_{j,k} \alpha_q \delta + O(\delta^2)
\end{aligned}
$$

which yields

$$\left[\frac{\partial \mathrm{E}\, \log\det(I + A^2)}{\partial \mathrm{Im}\, A_q}\right]_{j,k} = 4\mathrm{E}\,\mathrm{Im}\, \left[(I + A^2)^{-1}A\right]_{j,k} \alpha_q, \qquad j \neq k.$$

The gradient with respect to the diagonal elements is

$$\left[\frac{\partial \mathrm{E}\, \log\det(I + A^2)}{\partial A_q}\right]_{k,k} = 2\mathrm{E}\, \left[(I + A^2)^{-1}A\right]_{j,j} \alpha_q.$$

The third term in (C.1) has the same derivative as the second term.

For the fourth term, note that $A' - A = \sum_{q=1}^{Q} A_q \beta_q$ where $\beta_q = \alpha'_q - \alpha_q$. Therefore,

$$
\begin{aligned}
&\log\det(A + (e_j e_k^T + e_k e_j^T)\delta\beta_q)^2 \\
&\quad = \log\det(A(I + A^{-1}(e_j e_k^T + e_k e_j^T)\delta\beta_q))^2 \\
&\quad = \log\det(A(I + A^{-1}(e_j e_k^T + e_k e_j^T)\delta\beta_q)A(I + A^{-1}(e_j e_k^T + e_k e_j^T)\delta\beta_q)) \\
&\quad = \log\det A^2 + 2\mathrm{tr}\, \log(I + A^{-1}(e_j e_k^T + e_k e_j^T)\delta\beta_q) + O(\delta^2) \\
&\quad = \log\det A^2 + 2\mathrm{tr}\, \left[A^{-1}(e_j e_k^T + e_k e_j^T)\delta\beta_q\right] + O(\delta^2) \\
&\quad = \log\det A^2 + 2([A^{-1}]_{k,j} + [A^{-1}]_{j,k})\delta\beta_q \\
&\quad = \log\det A^2 + 4\mathrm{Re}\, \left[A^{-1}\right]_{j,k} \delta\beta_q.
\end{aligned}
$$

Hence,

$$\left[\frac{\partial \mathrm{E}\, \log\det(A' - A)^2}{\partial \mathrm{Re}\, A_q}\right]_{j,k} = 4\mathrm{E}\,\mathrm{Re}\, \left[A^{-1}\right]_{j,k} \beta_q, \qquad j \neq k.$$

For brevity, the computation of the derivatives with respect to the imaginary and diagonal components of A_q is omitted. The results are

$$\left[\frac{\partial \mathrm{E}\, \log\det(A' - A)^2}{\partial \mathrm{Im}\, A_q}\right]_{j,k} = 4\mathrm{E}\,\mathrm{Im}\, \left[A^{-1}\right]_{j,k} \beta_q, \qquad j \neq k$$

and

$$\left[\frac{\partial \mathrm{E}\, \log\det(A' - A)^2}{\partial A_q}\right]_{j,j} = 2\mathrm{E}\, \left[A^{-1}\right]_{j,j} \beta_q.$$

D Isotropic distribution achieves equality in (38)

The stationary points of the left-hand side of (38) can be obtained by considering the Lagrangian

$$\mathcal{L} = \frac{1}{M}\int dV dV' p_V(V) p_V(V') \log\det(V - V')(V - V')^* + \lambda(1 - \int dV p_V(V)) + \int dV \mu(V) p_V(V), \tag{D.1}$$

where λ is the Lagrange multiplier that enforces the constraint $\int dV p_V(V) = 1$, and $\mu(\cdot)$ is the Lagrange multiplier that enforces the constraint $p_V(\cdot) \geq 0$. In other words, $\mu(V) = 0$ whenever $p_V(V) \geq 0$ and $\mu(V) \neq 0$ otherwise. We require that $p_V(\cdot)$ be nonzero for all unitary V, and therefore $\mu(V) = 0$ and the Lagrangian becomes

$$\mathcal{L} = \frac{1}{M}\int dV dV' p_V(V) p_V(V') \log\det(V - V')(V - V')^* + \lambda(1 - \int dV p_V(V)).$$

Writing the first-order condition for maximization yields

$$\begin{aligned} \lambda &= \frac{2}{M}\int dV' p_V(V') \log\det(V - V')(V - V')^* \\ &= \frac{2}{M}\int dV' p_V(V') \log\det(I - V'V^*)(I - VV'^*) \end{aligned} \tag{D.2}$$

The distribution $p(V')$ is a stationary point if (D.2) holds for every fixed unitary V.

Suppose now that V in (D.2) is allowed to be random and is chosen according to the isotropic distribution. We take the expected value of (D.2) with respect to this distribution to obtain

$$\begin{aligned} \lambda &= \frac{2}{M}\mathrm{E}\int dV' p_V(V') \log\det(I - V'V^*)(I - VV'^*) \\ &= \frac{2}{M}\int dV' p_V(V') \mathrm{E}\ \log\det(I - V'V^*)(I - VV'^*) \\ &= \frac{2}{M}\int dV' p_V(V') \mathrm{E}\ \log\det(I - V'')(I - V''^*), \end{aligned} \tag{D.3}$$

where the expectation is over $V'' = V'V^*$ which, by properties of the isotropic distribution, is also isotropically distributed.

We simplify the expectation in (D.3) by diagonalizing $V'' = UDU^*$, where D is a diagonal matrix of eigenvalues denoted $d_1, \ldots, d_M$,

$$\mathrm{E}\ \log\det(I - V'')(I - V''^*) \quad = \quad \mathrm{E}\ \log\det(I - UDU^*)(I - UD^*U^*)$$

$$
\begin{aligned}
&= \mathrm{E}\ \log\det(I-D)(I-D^*) \\
&= \mathrm{E}\ \log\prod_{m=1}^{M}(1-d_m)(1-d_m^*) \\
&= \mathrm{E}\ \sum_{m=1}^{M}\log(1-d_m)(1-d_m^*) \\
&= M\mathrm{E}\ \log(1-d)(1-d^*) \\
&= M\mathrm{E}\ \log\left(4\sin^2\frac{\theta}{2}\right),
\end{aligned}
\tag{D.4}
$$

where $d = e^{i\theta}$ is an arbitrary eigenvalue of V'' and the expectation in (D.4) is over the distribution on θ. With an isotropic distribution on V'', θ has a uniform distribution on $[0, 2\pi)$, and hence,

$$
\begin{aligned}
\mathrm{E}\ \log\left(4\sin^2\frac{\theta}{2}\right) &= 2[\log 4 + \frac{1}{2\pi}\int_0^{2\pi} d\theta \log\sin^2\frac{\theta}{2}] \\
&= 2\log 4 + \frac{2}{\pi}\int_0^{\pi} d\theta \log\sin^2\theta \\
&= 2\log 4 + \frac{4}{\pi}\int_0^{\pi} d\theta \log\sin\theta \\
&= 2\log 4 - \frac{4}{\pi}\pi\log 2 \\
&= 0.
\end{aligned}
\tag{D.5}
$$

Equations (D.3), (D.4) and (D.5) imply that $\lambda = 0$ for every stationary point of the left-hand side of (38). Moreover, it is straightforward to use (D.3) to show that the value of (38) itself is also equal to λ and is therefore also zero at every stationary point. Because (38) is trivially bounded above by $\log 4$ and has no lower bound, any stationary point must be a maximum. We therefore conclude that

$$
\frac{1}{M}\mathrm{E}\ \log\det(V-V')(V-V')^* \le 0.
$$

Equality is achieved when $p_V(\cdot)$ is the isotropic distribution.

References

[1] G. J. Foschini, "Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas," *Bell Labs. Tech. J.*, vol. 1, no. 2, pp. 41–59, 1996.

[2] I. E. Telatar, "Capacity of multi-antenna gaussian channels," *Eur. Trans. Telecom.*, vol. 10, pp. 585–595, Nov. 1999.

[3] J.-C. Guey, M. P. Fitz, M. R. Bell, and W.-Y. Kuo, "Signal design for transmitter diversity wireless communication systems over Rayleigh fading channels," in *Proc. IEEE VTC*, pp. 136–140, 1996.

[4] S. M. Alamouti, "A simple transmitter diversity scheme for wireless communications," *IEEE J. Sel. Area Comm.*, pp. 1451–1458, Oct. 1998.

[5] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space-time codes for high data rate wireless communication: Performance criterion and code construction," *IEEE Trans. Info. Theory*, vol. 44, pp. 744–765, 1998.

[6] B. Hassibi and B. Hochwald, "High-rate codes that are linear in space and time," *submitted to IEEE Trans. Info. Theory*, 2000. Download available at `http://mars.bell-labs.com`.

[7] T. L. Marzetta, "Blast training: Estimating channel characteristics for high-capacity space-time wireless," in *Proc. 37th Annual Allerton Conference on Communications, Control, and Computing*, Sept. 22–24 1999.

[8] B. Hassibi and B. Hochwald, "How much training is needed in multiple-antenna wireless links?," *submitted to IEEE Trans. Info. Theory*, 2000. Download available at `http://mars.bell-labs.com`.

[9] T. L. Marzetta and B. M. Hochwald, "Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading," *IEEE Trans. Info. Theory*, vol. 45, pp. 139–157, 1999.

[10] B. M. Hochwald and T. L. Marzetta, "Unitary space-time modulation for multiple-antenna communication in Rayleigh flat-fading," *IEEE Trans. Info. Theory*, vol. 46, pp. 543–564, Mar. 2000.

[11] L. Zheng and D. Tse, "Packing spheres in the Grassman manifold: a geometric approach to the non-coherent multi-antenna channel," *submitted to IEEE Trans. Info. Theory*, 2000.

[12] B. Hassibi and T. L. Marzetta, "Block-fading channels and isotropically-random unitary inputs: The received signal density in closed-form," *Submitted to IEEE Trans. on Info. Thy.*, 2001.

[13] B. Hochwald, T. Marzetta, T. Richardson, W. Sweldens, and R. Urbanke, "Systematic design of unitary space-time constellations," *IEEE Trans. Info. Theory*, pp. 1962–1973, 2000.

[14] B. Hochwald and W. Sweldens, "Differential unitary space time modulation," *IEEE Trans. Comm.*, vol. 48, pp. 2041–2052, Dec. 2000. Download available at `http://mars.bell-labs.com`.

[15] B. Hughes, "Differential space-time modulation," *IEEE Trans. Info. Theory*, pp. 2567–2578, Nov. 2000.

[16] V. Tarokh and H. Jafarkhani, "A differential detection scheme for transmit diversity," *J. Sel. Area Comm.*, pp. 1169–1174, July 2000.

[17] Z. Liu, G. Giannakis, and B. Hughes, "Double differential space-time block coding for time-selective fading channels," in *Proc. IEEE WCNC*, 2000.

[18] J. G. Proakis, *Digital Communications*. McGraw-Hill, 2000.

[19] A. Shokrollahi, B. Hassibi, B. Hochwald, and W. Sweldens, "Representation theory for high-rate multiple-antenna code design," *submitted to IEEE Trans. Info. Theory*, 2000. Download available at `http://mars.bell-labs.com`.

[20] K. L. Clarkson, W. Sweldens, and A. Zheng, "Fast multiple antenna differential decoding," tech. rep., Bell Laboratories, Lucent Technologies, October 1999.

[21] B. Hughes, "Optimal space-time constellations from groups," *submitted to IEEE Trans. Info. Theory*, 2000.

[22] B. Hassibi and M. Khorrami, "Fully-diverse multi-antenna constellations and fixed-point-free Lie groups," *submitted to IEEE Trans. Info. Theory*, 2000. Download available at `http://mars.bell-labs.com`.

[23] A. Wittneben, "Basestation modulation diversity for digital simulcast," in *Proc. IEEE VTC*, pp. 848–853, 1991.

[24] G. D. Golden, G. J. Foschini, R. A. Valenzuela, and P. W. Wolniansky, "Detection algorithm and initial laboratory results using V-BLAST space-time communication architecture," *Electronic Letters*, vol. 35, pp. 14–16, Jan. 1999.

[25] G. J. Foschini, G. D. Golden, R. A. Valenzuela, and P. W. Wolniansky, "Simplified processing for high spectral efficiency wireless communication employing multi-element arrays," *J. Sel. Area Comm.*, vol. 17, pp. 1841–1852, Nov. 1999.

[26] B. Hassibi, "An efficient square-root algorithm for BLAST," *submitted to IEEE Trans. Sig. Proc.*, 1999. Download available at `http://mars.bell-labs.com`.

[27] U. Fincke and M. Pohst, "Improved methods for calculating vectors of short length in a lattice, including a complexity analysis," *Mathematics of Computation*, vol. 44, pp. 463–471, April 1985.

[28] M. O. Damen, A. Chkeif, and J.-C. Belfiore, "Lattice code decoder for space-time codes," *IEEE Comm. Let.*, pp. 161–163, May 2000.

[29] D. Warrier and U. Madhow, "Noncoherent communication in space and time," *submitted to IEEE Trans. Info. Theory*, 1999.

[30] B. Hassibi, T. Marzetta, and B. Hochwald, "Structured unitary space-time constellations," *submitted to IEEE Trans. Info. Theory*, 2000. Download available at `http://mars.bell-labs.com`.

[31] R. Horn and C. Johnson, *Topics in Matrix Analysis*. Cambridge: University Press, 1991.

[32] A. A. Sagle and R. E. Walde, *Introduction to Lie Groups and Lie Algebras*. Academic Press, 1986.

[33] L. Mirsky, *An Introduction to Linear Algebra*. Oxford: Clarendon Press, 1955.

[34] F. Diele, L. Lopez, and R. Peluso, "The Cayley transform in the numerical solution of unitary differential systems," *Advances in Computational Mathematics*, pp. 317–334, 1998.

[35] A. Iserles, "Cayley-transform algorithms for differential equations in Lie groups," in *Proc. of the Workshop on Geometrical Methods and Computation*, July 24–26, 1999.

[36] M. O. Damen, K. Abed-Meraim, and M. S. Lemdani, "Further results on the sphere decoder algorithm," *Submitted to IEEE Trans. Info. Theory*, 2000.

[37] X. B. Liang and X. G. Xia, "Unitary signal constellations for differential space-time modulation with two transmit antennas: Parametric codes, optimal designs and bounds," *submitted IEEE Trans. Info. Theory*, 2001.

[38] C. Peel and A. L. Swindlehurst, "Performance of unitary space-time modulation in a continuously fading channel," in *Submitted to Int. Conf. Acoust., Speech, Sig. Proc.*, 2001. Download available at `http://www.ee.byu.edu/~peel/papers/index.html`.

[39] R. Muirhead, *Aspects of Multivariate Statistical Theory*. New York: J. Wiley and Sons, 1982.

[40] A. Edelman, *Eigenvalues and Condition Numbers of Random Matrices*. PhD Thesis, MIT, Dept. of Math., 1989.